Department of Communications and Networking

Resource Allocation for 5G Systems

Bikramjit Singh

Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 169/2021

Resource Allocation for 5G Systems

Bikramjit Singh

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Electrical Engineering, at a public examination held at the lecture hall U006 of the school on 14 December 2021 at 12:00.

Aalto University
School of Electrical Engineering
Department of Communications and Networking

Supervising professor
Professor Olav Tirkkonen, Aalto University, Finland

Thesis advisors
Dr. Konstantinos Koufus, University of Bristol, England
Dr. Zexian Li, Nokia Bell Labs, Finland

Preliminary examiners
Associate professor Gilberto Berardinelli, Aalborg University, Denmark
Professor Allen B. MacKenzie, Tennessee Technological University, U.S.A.

Opponent
Professor Hirley Alves, University of Oulu, Finland

Aalto University publication series
DOCTORAL DISSERTATIONS 169/2021

ISBN 978-952-64-0614-5 (printed)
ISBN 978-952-64-0615-2 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-64-0615-2

Unigrafia Oy
Helsinki 2021

Finland

NORDIC SWAN ECOLABEL
Printed matter
4041-0619

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Bikramjit Singh

Name of the doctoral dissertation
Resource Allocation for 5G Systems

Publisher School of Electrical Engineering

Unit Department of Communications and Networking

Series Aalto University publication series DOCTORAL DISSERTATIONS 169/2021

Field of research Communications Engineering

Manuscript submitted 18 February 2021 **Date of the defence** 14 December 2021

Permission for public defence granted (date) 22 April 2021 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

The 5th generation of cellular communications is driven by three broadly classified use cases, enhanced mobile broadband, massive machine type communication, and ultra-reliable low latency communication. These new use cases bring a new set of problems such as provisioning of extremely high user rates, ultra-high reliability, ultra-low latency, extremely low jitter, and extremely long battery life, to name a few. Therefore, in this thesis, we focus on lower protocol layers of radio access networks, designing technology components enabling communication confirming new extreme requirements introduced by 5th generation systems.

To increase the amount of spectrum available for an operator in a specific geographic region, we discuss inter-operator spectrum sharing. The operators employ non-cooperative game-theoretic mechanisms to enable sharing without revealing their RAN-related information, e.g., optimization targets and traffic load to each other. The operators are independent and motivated by self-interest. However, if the sharing persists over a long time, non-cooperative operators may develop a degree of cooperation to maximize their self-interest by employing repeated games. Rational operators tend to cooperate due to fear of repercussions if they focus on optimizing long-term instead of short-term gains. Such cooperation emerges from individuals' greedy decisions. Simulation results show that the spectrum sharing based on repeated games provides better network utility than default spectrum allocations in the scenarios like limited spectrum pool or mutual renting.

In contrast to spectrum allocation, which focuses on the rate improvement for most users in a mobile broadband scenario, the other half of the thesis pivots to providing extremely reliable service, which induces the need to concentrate on the extremely deficient users in a service area. To improve critical communication, we consider allocation designs that boost rate availability and reliability. The availability targets a specific rate that a cell can provide at probable locations. We inspect multi-hop mechanisms for extreme availability, focusing on time scheduling weights, power allocation, and precoding weights. Inter-cell interference can limit availability at the cell edge, and thus we consider coordination among cells in limiting the interference. Furthermore, we consider allocation designs for minimum rate improvement using transmission repetitions for low latency contention-based access services. Simulation results show that the employed techniques significantly improve 5G ultra-reliable, low latency services.

Keywords 5G, Spectrum Sharing, Ultra-Reliable Low Latency Communication

ISBN (printed) 978-952-64-0614-5 **ISBN (pdf)** 978-952-64-0615-2

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2021

Pages 139 **urn** http://urn.fi/URN:ISBN:978-952-64-0615-2

Preface

The research work for this doctoral thesis has been carried out at the Department of Communications and Networking (COMNET) of Aalto University School of Electrical Engineering, Espoo, Finland. The work has been funded in parts by European Commission under project METIS, Nokia Bell Labs, Business Finland under project NSN WIVE, and the Academy of Finland under project CoMaCo.

First of all, I would like to thank my supervisor Professor Olav Tirkkonen for giving me this opportunity and for his continuous support and encouragement throughout my work. I am very grateful for his guidance and insights. I would like to extend my gratitude to Mikko Uusitalo for giving me the opportunity to work and guidance in a collaborative research project with Nokia Bell Labs. I would also like to thank my advisers Konstantinos Koufos and Zexian Li, and colleague Alexey Shapin who gave me day-to-day support and brought innovative ideas during discussions.

I would like to thank my co-authors for their contribution to the research work. I am also grateful to my colleagues and all the personnel at Aalto University School of Electrical Engineering, Espoo, Nokia Bell Labs, Espoo, and Ericsson, Jorvas, for their support and an exciting working environment.

I owe a special thanks to my family for all forms of support.

Espoo, November 12, 2021,

Bikramjit Singh

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I B. Singh, K. Koufos, O. Tirkkonen and R. Berry. Co-primary inter-operator spectrum sharing over a limited spectrum pool using repeated games. In *IEEE International Conference on Communications (ICC)*, pp. 1494-1495, June 2015.

II B. Singh, K. Koufos, O. Tirkkonen and R. Jäntti. Repeated spectrum sharing games in multi-operator heterogeneous networks. In *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, pp. 221-228, September 2015.

III B. Singh, S. Hailu, K. Koufos, A. A. Dowhuszko, O. Tirkkonen, R. Jäntti and R. Berry. Coordination protocol for inter-operator spectrum sharing in co-primary 5G small cell networks. In *IEEE Communications Magazine*, vol. 53, no. 7, pp. 34-40, July 2015.

IV B. Singh, Z. Li, O. Tirkkonen, M. A. Uusitalo and P. Mogensen. Ultra-reliable communication in a factory environment for 5G wireless networks: Link level and deployment study. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC) Workshop*, pp. 1-5, September 2016.

V B. Singh, O. Tirkkonen, Z. Li, M. A. Uusitalo and R. Wichman. Selective multi-hop relaying for ultra-reliable communication in a factory environment. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, pp. 1-6, September 2016.

VI B. Singh, O. Tirkkonen, Z. Li and M. A. Uusitalo. Interference coordination in

ultra-reliable and low latency communication networks. In *European Conference on Networks and Communications (EuCNC)*, pp. 251-255, June 2018.

VII B. Singh, O. Tirkkonen, Z. Li and M. A. Uusitalo. Contention-Based Access for ultra-reliable low latency uplink transmissions. *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 182-185, April 2018.

Author's Contribution

Publication I: "Co-primary inter-operator spectrum sharing over a limited spectrum pool using repeated games"

The author of this thesis contributed to formulating the idea, deriving the results, and writing this article.

Publication II: "Repeated spectrum sharing games in multi-operator heterogeneous networks"

The author of this thesis had a leading role in formulating the idea, deriving the results, and writing this article.

Publication III: "Coordination protocol for inter-operator spectrum sharing in co-primary 5G small cell networks"

The author of this thesis contributed to formulating the idea, deriving the results, and writing of article focusing on the sections pertaining to long-term periodicity and the combined coordination protocol.

Publication IV: "Ultra-reliable communication in a factory environment for 5G wireless networks: Link level and deployment study"

The author of this thesis had a leading role in formulating the idea, deriving the results, and writing this article.

Publication V: "Selective multi-hop relaying for ultra-reliable communication in a factory environment"

The author of this thesis had a leading role in formulating the idea, deriving the results, and writing this article.

Publication VI: "Interference coordination in ultra-reliable and low latency communication networks"

The author of this thesis had a leading role in formulating the idea, deriving the results, and writing this article.

Publication VII: "Contention-Based Access for ultra-reliable low latency uplink transmissions"

The author of this thesis had a leading role in formulating the idea, deriving the results, and writing this article.

List of Abbreviations and Symbols

Abbreviations

1G	1rd Generation
2G	2rd Generation
3G	3rd Generation
3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
ACK	Acknowledgment
ARQ	Automatic Repeat Request
BIM	Background Interference Matrix
BER	Bit Error Rate
BSR	Buffer Status Report
BLER	Block Error Rate
BW	Bandwidth
CB	Code Block
CBG	Code-Block Group
CC	Component Carrier
CCA	Clear Channel Assessment
CDF	Cumulative Density Function
CG	Configured Grant
CORESET	Control Resource Set
CQI	Channel Quality Indicator
CSI	Channel State Information
DC	Dual Connectivity
DCI	Downlink Control Information
e2e	end-to-end
eMBB	enhanced Mobile Broadband

FR	Frequency Range
GF	Grant-Free
gNB	Next generation nodeB
GPRS	General Packet Radio Service
GBR	Guaranteed Bit Rate
HARQ	Hybrid Automatic Repeat Request
HetNet	Heterogeneous Network
IEEE	Institute of Electrical and Electronics Engineers
IIoT	Industrial Internet of Things
IMT	International Mobile Telecommunications
IP	Internet Protocol
ISM	Industrial, Scientific, and Medical
KPI	Key Performance Indicator
LBT	Listen Before Talk
LC	Logical Channel
LDPC	Low-Density Parity Check
LSP	Limited Spectrum Pool
LTE	Long Term Evolution
M2M	Machine-to-Machine
MAC	Medium Access Control
MBB	Mobile Broadband
MCS	Modulation Coding Scheme
MIMO	Multiple-Input and Multiple-Output
MMF	Max-Min Fair
mMTC	massive Machine Type Communication
mmWave	millimeter Wave
MTC	Machine Type Communications
Multi-TRP	Multiple Transmission and Reception Points
MR	Mutual Renting
NB-IoT	Narrowband-Internet of Things
NE	Nash Equilibrium
NR	New Radio
NR-U	NR Unlicensed
OFDM	Orthogonal Frequency-Division Multiplexing
PDCCH	Physical Downlink Control Channel
PDCP	Packet Data Convergence Protocol
PDF	Probability Density Function
PDSCH	Physical Downlink Shared Channel

PER	Packet Error Rate
PF	Proportional Fair
PHY	Physical Layer
PPP	Poisson Point Process
PUSCH	Physical Upink Shared Channel
QoS	Quality-of-Service
RB	Resource Block
RAN	Radio Access Network
RAT	Radio Access Technology
RLC	Radio Link Control
RN	Relay Node
RRC	Radio Resource Control
SA	Service and System Aspects
SCS	Subcarrier Spacing
SINR	Signal-to-Interference-plus-Noise Ratio
SMS	Short Message Service
SNR	Signal-to-Noise Ratio
SPS	Semi-Persistent Scheduling
SR	Scheduling Request
SVD	Singular Value Decomposition
TB	Transport Block
TBS	Transport Block Size
TDD	Time Division Duplex
TDRA	Time Domain Resource Allocation
TSN	Time-Sensitive Networking
TTI	Transmission Time Interval
UE	User Equipment
URLLC	Ultra-Reliable and Low Latency Communication
V2X	Vehicle-to-Everything
XR	Extended Reality

Symbols

$\mathbb{1}$	Identity matrix
α	Parameter weighted fairness
δ	Discount factor
γ	Signal-to-interference-plus-noise ratio
Γ	Transmission repetition or diversity
η	Path loss exponent
θ	Utility gain threshold
λ	Utility loss threshold
ρ	Rate observed over a unit time-period
κ	Number of antennas on a node
σ	Rayleigh distribution scale parameter
ς	Access slot number
ν	Noise standard deviation
τ	Time scheduling weight
ψ	Traffic intensity
ζ	Standard deviation of Normal distributed shadowing
$\mathcal{B}$	Bandwidth
BIM	Background interference matrix
d	Distance between the transmitting and receiving nodes
f	Probability density function of utility losses or gains
g	Instantaneous utility gain
h	Channel coefficient
$\mathbf{H}$	Channel matrix
I	Interference power level of a node/antenna
$\mathbf{I}$	Interference matrix
K	Number of resource blocks or component carriers
ℓ	Transmission latency
l	Instantaneous utility loss
L_d	Path loss at distance d
M	Numbers of antennas
N	Numbers of users
$\mathcal{N}$	Normal distribution
p	Transmission power of a node
P^{ask}	Probability of asking a favor
P^{col}	Probability of collisions
P^{grant}	Probability of granting a favor

P^{rel}	Reliability target
r	Transmission rate
$\mathcal{R}$	Rayleigh distribution
s	Strategy of a player
S	Strategy set of a player
t	Time
T	Latency budget
U	Utility of a player
v	Normal distributed variable
$\mathbf{W}$	Precoding matrix
x	User specific quantity
$\mathbf{X}$	Covariance matrix

1. Introduction to the Thesis

1.1 Motivation

Mobile data traffic continues to rise, and it is expected to reach 131 Exabytes per month by 2024 [1]. On top of this continually increasing mobile traffic, related to traditional human communication, completely new use cases, e.g., self-driving cars, industrial automation, smart grid, and high-end gaming features like augmented reality [2], will also grow and evolve. Fig. 1.1 depicts various services and applications that are and will be supported by mobile traffic in future wireless networks. For example, in the areas of intelligent driving, 100 million cellular-Vehicle-to-Everything (V2X) subscriptions are anticipated, and the figure is expected to be half a billion by 2028 [3], and the overall Machine-to-Machine (M2M) subscriptions are estimated at 70 billion [4].

Multiple studies [1, 3–6] present predictions on future mobile traffic. These works reflect a clear trend that wireless traffic continues to rise enormously. Moreover, current cellular technologies such as 3rd Generation (3G), 4th Generation (4G)/Long Term Evolution (LTE), or Wi-Fi technologies may not be sufficient to cater to the ever-growing requirements, nor to the demands of completely new future services with strict Quality-of-Service (QoS) [2, 7].

The generations of wireless cellular communication up to 4G are aimed at human-centric voice and data communication [8]. The 1st Generation (1G) was used to make voice calls, and networks were analog. The 2nd Generation (2G) employed digital communication had the additional Short Message Service (SMS), and limited data services using the General Packet Radio Service (GPRS) protocol. Later, 3G has enabled additional features, such as mobile internet access, video calls, and mobile TV. Driven by the needs of a Mobile Broadband (MBB) usage scenario of providing extremely high data rates, 4G LTE networks were designed to offer data and voice services over the Internet Protocol (IP).

Figure 1.1. IMT-2020 and beyond services [2].

5th Generation (5G) communication systems trial and non-standalone deployment have already begun, facilitating the expansion of International Mobile Telecommunications (IMT) beyond IMT-2000 and IMT-Advanced and supporting diverse usage scenarios and services. The usage scenarios are classified into three categories [2, 9]

- enhanced Mobile Broadband (eMBB),
- Ultra-Reliable Low Latency Communication (URLLC), and
- massive Machine Type Communications (mMTC).

The eMBB use case addresses human-centric needs for access to multi-media content, services, and data. Much-improved User Equipment (UE) data rates are desired compared to existing rates. In addition seamless coverage and good mobility are required for operating in a wide coverage area. With increasing UE density and service needs, hotspots with very high traffic capacity are foreseen to emerge [2, 9, 10].

URLLC is a new usage scenario that enables real-time control and automation of dynamic processes for vertical applications. Such services require very high reliability or low Block Error Rate (BLER) and often short latency down to millisecond level [2, 9–11].

The mMTC use case is embodied with massive number of low-cost connected devices that typically transmit a relatively low volume of delay tolerant data. In addition, devices are expected to have a very long battery life [2, 9, 10, 12]. Fig. 1.2 portrays various capabilities that are needed for fulfillment of the three discussed usage scenarios.

This thesis focuses on scenarios involving eMBB and URLLC services that demand spectrum efficiency, high reliability, and low latency. Radio Access Network (RAN) design and optimization is needed when addressing the requirements of use cases, and therefore plays a key role in 5G system realization. The two use cases can be differentiated based on reliability and latency requirements. The eMBB service is

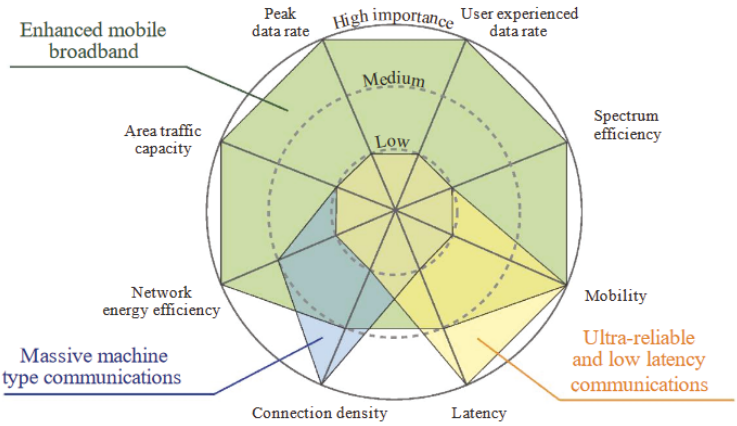

Figure 1.2. The salient features in different usage scenarios [2]

characterized by high data rates and a tolerance for a high BLER, e.g., 10 %, and a large latency budget. To improve eMBB service, techniques like massive Multiple-Input and Multiple-Output (MIMO), additional spectrum use, e.g., millimeter frequencies, spectrum sharing with licensed or license-exempt spectrum, and Code-Block Group (CBG) based transmissions can be applied [15]. URLLC services require an extremely high reliability of the order of 99.999 % and a latency budget down to 0.5-1 milliseconds. Examples of techniques [15] facilitating high reliability are the utilization of low Modulation and Coding Scheme (MCS) or repetition coding. Low latency can be enabled with mini-slot transmission or Semi-Persistent Scheduling (SPS) based allocation.

1.1.1 Research questions

The focus of the thesis is on eMBB and URLLC services. For eMBB, we consider a multi-operator setting, which is relevant for 5G deployment, especially in higher frequencies. There are two fundamental reasons for this: first, higher frequencies suffer from significant path loss, and second, Physical (PHY) layer operation in higher frequencies is based on narrow beams. This favors spectrum sharing or coordination among operators as inter-cell interference may not be as strong as in lower frequencies. Thus, residual inter-cell interference can be curtailed with proper coordination among sharing operators or cells. For URLLC services, we target existing Medium Access Control (MAC) designs, which have conventionally been designed to deliver, e.g., MBB services in cellular (dedicated access) and Wi-Fi networks (contention-based access). We strive to enhance these mechanisms to deliver URLLC services that exhibit QoS requirements diametrical to MBB services for cellular operation.

Therefore, the two key research questions which we intend to explore in this thesis

are:

1. To what extent can neighboring competitive operators coordinate and share spectrum to improve eMBB services in their geographical area?

2. How can we ensure that MAC mechanisms, whether based on dedicated access, or contention, can fulfill extreme reliability, availability, and low latency requirements to aid URLLC applications?

The detailed prior-art study, their shortcomings, and research problems are discussed in specific chapters related to eMBB and URLLC scenarios, i.e., in Chapter 3 and 4.

1.2 Research methodology

The studies are carried out by a combination of analytical and simulation approaches. The overall working methodology can be summarized as follows:

- Problem scenario: The main problems are identified by analyzing the literature, and ongoing 3rd Generation Partnership Project (3GPP) efforts in driving study and work items for next generation networks. In this thesis, the issues centered around spectrum deficiency for eMBB services and reliability, availability, and latency access inadequacy for URLLC services are discussed.

- Analysis and problem solving: Next, the relevant research questions in the scenarios are formulated in terms of analytical models. Based on these formulations, solutions are proposed to address the identified problem and possible benefits are identified. A brief overview of modeling PHY and MAC layers, characterizing RAN performance is presented in Chapter 2. Detailed analytical modeling for the considered scenarios is discussed in Chapter 3 and 4, focusing on the spectrum sharing and URLLC, respectively.

- Proposal validation: The analytical model and proposed solutions were tested employing Monte Carlo system-level simulations. A time-based simulator is developed in MATLAB. In the simulator, values of Key Performance Indicators (KPIs), such as bit rate, availability, reliability, and latency, are collected. System-level simulation is carried out in scenarios representing typical cellular network settings accommodating multiple UEs. Results are usually plotted in Cumulative Distribution Functions (CDFs).

- Analysis: The simulation plots are compared with analytical plots to verify the accuracy of models. Some analytical models are simplified by considering approximations suitable for targeted scenarios. Therefore, deviations are

expected when models are applied to non-compatible scenarios, and such analyses have also been discussed. Further, the proposed solutions are compared with baseline models or scenarios to exhibit benefits. At the end of the entire procedure, the discussed solutions with simulation results and analysis are presented in peer-reviewed scientific publications.

1.3 Contribution

The thesis contribution is divided into two parts. In the first contribution, we address the resource allocation issue for eMBB traffic demand. Network operators are interested in providing increased data rates to their UEs, and therefore, they are interested in sharing their spectrum. We considered a pragmatic setting where operators do not reveal any critical RAN-related information. As a consequence, we consider non-cooperative sharing strategies. Strategies are modeled with game theory, where operators play repeated games by exchanging spectrum favors and maximize their self-interest.

The second part of the thesis focuses on resource allocation for URLLC services. These services require extremely high availability or reliability of the order of 99.999-99.9999 $\%$ within a latency bound of 0.5-1 ms. We target both grant-based (dedicated scheduling) and Grant-Free (GF) data transmissions with low latency access. In grant-based access, resource optimization techniques are devised that improve rates of deficient UEs and thus improve availability. In scenarios with GF access, reliability may suffer due to shared resource use. Hence, we employ and investigate protocols influenced with repetition coding to benefit these UEs.

1.4 Structure

The remainder of the thesis is organized as follows. Chapter 2 provides a brief overview of the 5G system and its physical resources. Chapter 3 discusses a game-theoretic framework for non-cooperative co-primary spectrum sharing. A non-cooperative model is discussed to implement the coordination protocol among operators by exchanging spectrum-based favors. Next, Chapter 4 discusses URLLC resource allocation strategies. It presents optimization techniques that improve the availability and reliability of URLLC services and are supported with performance analysis. Conclusions and future work are discussed in the last chapter.

2. 5th Generation System and Resource Overview

2.1 5th generation system

2.1.1 Rationale

The progression from 4G to 5G will enable expansion of device connectivity and, at the same time, supporting devices with ultra-high communication speed. 5G will enable the telecommunication industry to expand toward a “vertical" delivery model. Services are tailored to specific industry sectors, in contrast to the “horizontal" service delivery model, where services, predominantly MBB, are defined independently of their consumers. In 5G New Radio (NR) new technological features have been introduced, which provide capacity, latency, and reliability improvements, as well as unlimited connectivity. These features can help improve traditional human-centric services such as mobile broadband and enable two new use cases - URLLC and mMTC. Apart from these three fundamental use cases, 5G features can facilitate a mix of use cases. One example is the support for Extended Reality (XR), a combination of URLLC and eMBB services, i.e., XR service is characterized as high capacity and reliable service with bounded latency [13]. Hence, the support for URLLC and mMTC in 5G enable a vertical delivery model, which can revolutionize the industrial value chain, enabling smart manufacturing, and paving the way for Industry 4.0 [14] or Industrial Internet of Things (IIoT).

Next, we briefly describe 5G NR and some of its capacity, reliability, and latency improvement features. These features are standardized or discussed in 3GPP aiding eMBB and URLLC use cases, which are emphasized in this thesis.

2.1.2 Brief description

5G, the next generation standard for wireless communications, is currently being standardized by 3GPP with a completely new air interface 5G NR. The first standard related to 5G was Release 15 [15]. It laid the foundation for 5G, providing faster speed, higher capacity, reliable communication, and shorter latency. Enhancements are set to follow in ongoing NR Release 17 and 18 work [15].

The most notable difference between NR and its predecessor LTE is numerology, i.e., Subcarrier Spacing (SCS) and symbol length. LTE has only one type of SCS, 15 kHz with subframe or Transmission Time Interval (TTI) size 1 ms, the minimum scheduling time granularity. While in 5G, a slot is used as a scheduling unit with options to choose higher SCSs [16], with reduced slot size. For example, with 240 kHz SCS, slot size is 62.5 µs in contrast to the 1 ms of LTE, each containing 14 Orthogonal Frequency-Division Multiplexing (OFDM) symbols. This drastically reduces latency and makes 5G communication significantly more responsive.

5G networks are set to operate in bands with higher carrier frequency in addition to more conventional Bandwidth (BW) alternatives [17]. The spectrum band terminology of 5G is as follows:

- Frequency Range (FR) 1: It has a frequency range consisting of low-band (450 MHz to 1 GHz or sub-1 GHz band) and mid-band (1 - 6 or 7 GHz). It is also known as sub-6 or sub-7 GHz band. FR 1 can be used in network deployment with 2G to 5G.

- FR 2: It is referred to a spectrum range higher than 24 GHz, which can go up to 71 GHz. These frequencies are part of Millimeter Wave (mmWave) band, which is the band of spectrum between 30 GHz and 300 GHz with wavelengths are of the order of $\sim 1 - 10$ mm. FR 2 (sub-100 GHz mmWave) is only used for 5G network deployment.

5G is also expected to operate in license-exempt frequencies in the 5 GHz band and the new 6 GHz band. The related technology component suite is known as NR Unlicensed (NR-U) [18, 19]. Hence, with supplemented BW and techniques like massive MIMO, 5G capacity will greatly increase.

5G will play an essential role in factory automation and Industry 4.0. This requires data transmission in a timely manner under extremely tight reliability, latency, and synchronization constraints. Such requirements can be fulfilled with Time-Sensitive Networking (TSN), which defines a set of Institute of Electrical and Electronics Engineers (IEEE) standards primarily for Ethernet. However, TSN is not desirable for future industries with manufacturing automation and mobility due to its wired na-

ture. 3GPP standardization RAN and Service and System Aspects (SA) groups are currently working to support TSN integration and URLLC for 5G [20], which will enable greater and better adoption of industrial automation. Stringent requirements such as 1 µs end-to-end (e2e) time synchronization accuracy, high reliability of 99.9999 %, or low air interface latency of 1 ms are some of the examples of requirements that have emerged in this standardization process [21, 22]. In the following, we present some notable features that can significantly benefit eMBB and URLLC services.

2.1.3 Enhanced mobile broadband scheduling features

5G NR aims to deliver eMBB services supporting ever-increasing end-user data rate and system capacity. To achieve this target, 5G has adopted many new features.

- *Slot-aggregation*

  LTE Time Division Duplex (TDD) has seven pre-defined configurations for downlink and uplink allocation in a radio frame [23]. For example, for traffic with a downlink-heavy pattern, Configuration 5 is the most suitable with eight downlink subframes, and for uplink-heavy traffic, Configuration 0 allocates six uplink subframes per radio frame. NR allows superior capacity by aggregating all the slots available in a radio frame. For example, for downlink-heavy traffic, a network can aggregate nine downlink slots with Format 0 and the last slot composed of both downlink and uplink symbols, e.g., configured with Format 33 in a 10 ms radio frame with 15 kHz SCS. The uplink symbols are used to transmit Hybrid Automatic Repeat Request (HARQ)-Acknowledgment (ACK) feedback for downlink transmissions. To serve uplink-heavy traffic, a network can aggregate first slot, e.g., configured with Format 45, containing both downlink and uplink symbols and remaining nine uplink slots with Format 1 in a radio frame. The downlink symbols are used to transmit Downlink Control Informations (DCIs) for uplink grants [24].

- *Carrier aggregation*

  LTE carriers have a maximum BW of 20 MHz that can be aggregated to achieve a channel BW of 100 MHz in LTE-Advanced, or 640 MHz with 32 Component Carriers (CCs) in LTE-Advanced Pro. NR, in contrast, has a maximum carrier BW of 1000 MHz in FR 1 and 400 MHz in FR 2. NR carriers can be aggregated up to a maximum BW of 2.7 GHz, and the maximum number of CCs that can be concatenated is 14 [25, 26].

- *Shared channel coding*

  NR employs Low-Density Parity Check (LDPC) codes for shared channel processing, while Turbo codes were used in LTE. LDPC codes provide better

performance at high code rate when code blocks are large. This leads to performance improvement in eMBB use cases [27, 28].

- *Control resource set*
  NR has introduced the concept of a Control Resource Set (CORESET) [16] to transmit control information, similar to the LTE control region [29]. CORESET can be flexibly configured within a BW part and thus is more spectrally efficient, which means more resources can be utilized for data transmission. In LTE, the control region is spread across the whole BW.

2.1.4 Ultra-reliable, low latency communication scheduling features

5G offers numerous solutions that help to provide reliable communication within a given latency budget. The solutions can be either purely implementation issues, such as scheduling strategies defined by operators, or new features incorporated in NR standardization [16–18, 24–26, 30, 31, 33]. Some of these features are discussed as follows.

A. High reliability features

- *Control channel coding*
  NR utilizes Polar codes for control channel processing, unlike LTE, which employs Convolutional codes. The polar codes are advantageous for small Transport Block (TB) sizes and thus increase reliability performance [27, 34].

- *Code-Bock Group*
  In LTE and NR, a large TB can be split into multiple Code Blocks (CBs). NR can group multiple CBs into a CBG and retransmit CBG upon receiving negative ACK. This approach has the advantage of reliable transmission of data with large Transport Block Sizes (TBSs) when there are limited resources. For example, a transmission with a certain BLER has a retransmission probability equivalent to its BLER target, and if the large TBS is split into smaller CBs, then retransmission would not require the transmission of the whole TBS [31].

- *Repetitions*
  5G enables data channel transmission with multiple repetitions for robust delivery. The repetitions can be selected statically with Radio Resource Control (RRC) configuration or dynamically with DCI by indicating the appropriate Time Domain Resource Allocation (TDRA) field [31]. LTE-Machine Type Communications (MTC) [29] also incorporated this feature but only for low-power devices with infrequent communication [35].

- *Channel quality indicator and modulation coding scheme tables*
  A new Channel Quality Indicator (CQI) and MCS tables are introduced for a BLER target of 10^{-5} [31] to support URLLC services.

- *Duplication over the packet data convergence protocol layer*
  Akin to LTE, NR allows splitting of bearers at the Packet Data Convergence Protocol (PDCP) layer and transmission via two radio paths in Dual Connectivity (DC). However, NR goes one step further by duplicating the data in the PDCP layer and thus improving transmission reliability [30, 36].

B. Low latency features

- *Non-slot based transmission*
  NR allows scheduling over mini-slots or sub-slots, which can be of length of few OFDM symbols. In the downlink, the minimum length of the data channel can be allocated with two symbols, whereas for uplink, it further diminishes to one symbol [16]. By reducing transmission size, latency can be curtailed. In addition, multiple TBs can be packed within the same slot and thus can be transmitted within a small latency budget.

- *Out-of-order transmission*
  LTE and NR Release 15 have a scheduling restriction which states that a Next Generation NodeB (gNB) allocates Physical Downlink Shared Channels (PDSCHs) or Physical Uplink Shared Channels (PUSCHs) according to the order of their DCIs to a UE. The restriction also applies to feedback resources for different HARQ processes which must be allocated according to their PDSCHs order. This restriction is removed partially in NR Release 16, where out-of-order transmissions, including overlapping transmissions, can be allowed with Multiple Transmission and Reception Points (Multi-TRP) for PDSCH transmissions [31].

- *Configured grant*
  NR introduces an uplink Configured Grant (CG) where a UE can autonomously transmit its uplink data on a pre-configured PUSCH allocation periodically. It is similar to LTE SPS utilized for voice communication but more reliable, e.g., a CG transmission can be applied with multiple repetitions [31, 32]. The benefit of CG allocation is that no latency budget is spent for signalings such as Scheduling Request (SR) and scheduling grant DCI, and therefore, the transmissions incur low latency as data is transmitted readily on CG's occasions.

- *Multiple configured grants*
  5G enables a UE to transmit over multiple CGs, allowing UEs to transmit mul-

tiple periodic services. Another usage is to deter misalignment errors due to fluctuating traffic. With a single CG allocation only, if a TB arrives late in a CG period, then the UE must wait until the next period for the transmission. On the other hand, if a UE is allocated with multiple CGs with different time offsets, the UE can transmit its TB in the nearest CG resource. In the uplink, the UE can have a maximum of 12 CGs, and in the downlink, at least 8 SPS grants can be allowed [37].

- *Puncturing*
  For an urgent URLLC transmission, gNB can interrupt an ongoing less urgent eMBB transmission, and puncture its resource, using it for a small URLLC transmission instead. This may degrade eMBB reception reliability. In order to reduce negative impacts, a puncturing indicator can be sent to the victim UE to exclude the punctured part from data decoding [24].

- *User processing time*
  In LTE, when a UE receives a PDSCH transmission in subframe n, it is required to transmit HARQ-ACK earliest in subframe in $n+3$ subframe. Also, a transmission of an uplink grant on DCI in subframe n cannot allocate PUSCH earlier than in subframe $n+3$ [29]. However, NR considerably reduces these limits derived from UE processing time. Instead NR introduces UE capabilities, e.g., Capability 1 and 2, which define the maximum processing time down to a few symbols. Thus, physical shared channels and feedback channels can be allocated with a shorter span in NR depending on UE's capability [31].

- *Shared channel coding*
  NR utilizes LDPC codes for data channels which are parallel in nature, while the Turbo codes used in LTE are serial. This enables LDPC to support latency-sensitive applications [27, 38].

2.2 System modeling

In this thesis, to estimate eMBB and URLLC performance, a model is required to test existing features and their enhancements. In this section, we cover abstract models of the cellular system. In system modeling, only the performance of lower layers of the RAN protocol stack is considered, namely the PHY and MAC layers.

2.2.1 Physical layer modeling

The Physical Layer serves as an interface between the MAC layer and the physical environment. PHY parameterization determines transmission behavior in cellular communication, e.g., its latency, reliability, and capacity, by selecting or configuring parameters related to, for instance, MCS, repetitions, Channel State Information (CSI) feedback, frame structure formats, and signaling formats. PHY parameters can depend on transmission requirements, infrastructure, physical channel proprieties, and traffic. Hence, PHY modeling furnishes tools that help to design PHY layer parameters subject to the transmission medium impact and traffic needs.

A. Physical environment

In cellular communication, information is carried using radio waves. The transmitter generates a carrier wave using an oscillator, and an information signal is pressed upon the carrier wave with an appropriate modulating scheme. The propagation medium distorts the transmitted signal. The impact of the medium is characterized in terms of distance-dependent path loss, shadowing, fast fading, multipath propagation, and Doppler effects. In this thesis, three types of channel effects have been considered:

- Large scale variations
  - Path loss — a deterministic long-term loss attributed to the distance d between gNB and UE. Path loss is modeled as $L_d = L_{d_o} + 10\ \eta \log_{10}\left(\frac{d}{d_o}\right)$ [dB] where L_{d_o} [dB] is path loss at reference distance d_o, and η is the path loss exponent.
  - Shadowing — a loss caused by large obstacles which produce scattering of transmission. It is modeled as a log-normal distribution $L_s \sim \mathcal{N}\left(0, \zeta^2\right)$. Hence, when shadowing power levels are measured in dB, they follow a Gaussian distribution with a mean zero and variance ζ^2 in dBs.
- Small scale (fast) fading — a loss caused by change of the transmission amplitude and phase over short periods of time, related to the movement of the UE of a carrier frequency wavelength. In this thesis, fast fading is modeled as Rayleigh fading, which can be understood to arise from a set of reflected waves with no dominant component present. Rayleigh fading is modeled as the absolute value of a zero-mean, unit variance circular complex Gaussian random variable, i.e., $a_f \sim \mathcal{R}\left(\frac{\sigma}{\sqrt{2}}\right)$ is Rayleigh distributed with scale parameter[¶] $\frac{\sigma}{\sqrt{2}}$ if $h_f \sim \mathcal{CN}\left(0, \sigma^2\right)$ is a complex normal random variable with zero-mean and

[¶]Correction: In [PIV], σ^2 is referred to as a variance of Rayleigh distribution, which is imprecise. It is the variance of the underlying complex Gaussian distribution. The scale parameter of Rayleigh

variance σ^2, i.e., $h = v_1 + \iota v_2$, where $v_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{2}\right)$ and $v_2 \sim \mathcal{N}\left(0, \frac{\sigma^2}{2}\right)$ are independent normal random variables. In system modeling, we are interested in received powers, i.e., channel gains. When the amplitude of a signal is Rayleigh distributed with scale parameter $\sigma/\sqrt{2}$, the channel gain is exponentially distributed with scale parameter $1/\sigma^2$. The mean of the exponentially distributed variable is then σ^2, and the variance is σ^4.

By combining all losses, the complex channel coefficient is $h = \sqrt{10^{\frac{-L_d - L_s}{10}}} h_f$, where $h_f \sim \mathcal{CN}(0, 1)$, while L_d and L_s are the distance dependent and shadowing losses in dB. The channel amplitude is then $a = \sqrt{10^{\frac{-L_d - L_s}{10}}} a_f$, with $a_f \sim \mathcal{R}\left(1/\sqrt{2}\right)$, and the channel gain is $g = 10^{\frac{-L_d - L_s}{10}} g_f$, where g_f is exponentially distributed with mean 1.

B. Signal-to-interference-plus-noise ratio

For eMBB capacity and URLLC availability modeling, we consider a function of transmission rate as a performance metric. The rate r is formulated in terms of Shannon capacity [39], which is a measure of the maximum amount of information that can be transmitted over a channel with BW $\mathcal{B}$ and Signal-to-Interference-plus-Noise Ratio (SINR) γ:

$$r = \mathcal{B} \log_2 (1 + \gamma) . \tag{2.1}$$

The SINR changes according to the path gain of the medium and thus requires recurring input from the PHY model to adapt to the changes. For eMBB modeling, the SINR averaged over fast fading is considered when determining communication rate. This provides a stable input to strategies set to improve eMBB cellular performance. For URLLC modeling, the instantaneous SINR is considered for PHY modeling due to the reliable communication requirement. In case of κ-stream MIMO communication between the base station with M_B antennas and UEs with M_U antennas, the rate is estimated as[‡]

$$r = \mathcal{B} \log_2 \det \left(\mathbb{1} + \mathbf{W}\ \mathbf{H}\ \mathbf{H}^{\mathrm{H}} \mathbf{W}^{\mathrm{H}} \mathbf{X}^{-1} p\right)$$

where $\mathbf{H}$ is the $M_B \times M_U$ channel matrix, $\mathbf{W}$ is the $\kappa \times M_B$ precoding matrix at gNB, $\mathbf{X}$ is the noise-plus-interference covariance matrix, $\mathbb{1}$ is the κ x κ identity matrix, and p is the transmission power of gNB. The noise-plus-interference covariance matrix is $\mathbf{X} = \nu^2 \mathbb{1} + \mathbf{I}$, where ν^2 is the noise variance, and $\mathbf{I}$ is the covariance matrix of the aggregate inter-cell interference inflicted on the UE. It is assumed that the gNB possesses knowledge of CSI and can apply the Singular Value Decomposition (SVD) technique to decompose MIMO channels to κ orthogonal eigenchannels.

distribution is σ [40]. The variance is $\frac{4-\pi}{2}\sigma$.

[‡]Notation: The complex conjugate transpose (or Hermitian) operation on a matrix is denoted by $(\cdot)^{\mathrm{H}}$.

2.2.2 Medium access control layer modeling

In wireless networks, the MAC protocol authorizes several UEs to transmit over the wireless medium and to share the medium's capacity. The primary functions of the MAC layer are mapping of Logical Channel (LC) to TB, LC prioritization, error handling via HARQ, multiplexing/demultiplexing of MAC service data units, multiplexing of the traffic of different UEs sharing the same channel, as well as controlling discontinuous reception, resource requests, uplink timing, and power headroom reporting [41]. MAC protocols belong to two categories:

- Grant-based (scheduled or contention-free) allocation,
  - Dynamically scheduled allocation, e.g., dynamically scheduled PDSCH or PUSCH transmissions,
  - Semi-persistently scheduled allocation, e.g., SPS or CG transmissions,
- Grant-free (contention-based) allocation,
  - With channel sensing algorithms, e.g., Clear Channel Assessment (CCA), Listen Before Talk (LBT) [42],
  - Without channel sensing algorithms.

In this thesis, the studies on spectrum allocation and URLLC services consider grant-based scheduling and contention-based allocation, but without channel sensing.

A. Grant-based allocation

To measure downlink eMBB and URLLC availability performance, we have considered a grant-based model. A deployment scenario is considered with multiple cells or gNBs where, at a particular time instant, the distribution of UEs is modeled by a Poisson Point Process (PPP) with a mean load N UEs per cell. These cells can belong to different operators and thus may or may not agree to share part of their BW resources depending on sharing protocols in use. Within each cell and during each scheduling time instant, a single UE is served with full-buffer downlink traffic over the whole cell BW. A gNB constructs a network utility U, which measures the QoS of its served UEs and distributes resources, for instance, time scheduling weights, among its UEs, to maximize utility. The gNB thus solves the optimization problem

$$\begin{aligned} \underset{\tau_n}{\text{Maximize}}: \quad & U(\tau_1, \tau_2, \ldots, \tau_N), \\ \text{Subject to}: \quad & \sum_{n=1}^{N} \tau_n = 1, \\ & t_n \geq 0, \, \forall \{n\}. \end{aligned} \tag{2.2}$$

Utility

The utility function represents network performance or QoS [43]. It is formulated as a strictly non-decreasing and concave function of the system parameters. Utility-based approaches can be adopted to quantify radio resource allocation problems in wireless communication. The function describing system-wide utility and welfare functions studied in economic sciences bears the same characteristics.

Utility functions have been used to model various performance parameters, for instance, data traffic (Shannon capacity) [44], BW allocation [45], multi-UE diversity [46], scheduling/delay-tolerant traffic [47], Signal-to-Noise Ratio (SNR)/SINR improvement, BW pricing applications [48], fairness [49, 50], Bit Error Rate (BER), and energy efficiency [51]. In this thesis, the focus is on studying fairness in eMBB systems where such services are often described with capacity-based utility functions, depicted in Fig. 2.1.

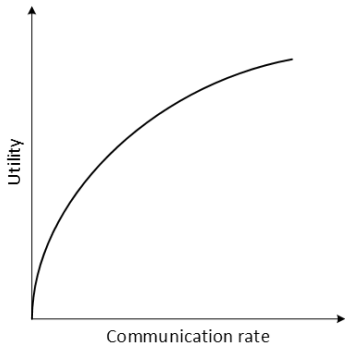

Figure 2.1. Example of utility function behavior.

Consider a cell with a load of N UEs. The base station or gNB quantifies the UE preference or experience in terms of resource allocation. The associated utility function with UE preference can be expressed as $U\left(x_n\left(t\right)\right)$, where $x_n\left(t\right)$ denotes UE specific quantity, e.g., an experienced rate of n-th UE at time t, and U is a function that describes UE satisfaction level. Utility function $U\left(x_n\left(t\right)\right)$ is an increasing and strictly concave function representing decreasing additional benefits with increasing resource allocation. From UE's satisfaction perspective, gNB is interested in optimizing a resource allocation, e.g., carrier allocation or time weights denoted by strategy s from resource constraint set $\mathcal{S}$, which maximizes long-term expected aggregated utility,

$$\max_{s\in\mathcal{S}} \lim_{\mathcal{T}\to\infty} \frac{1}{\mathcal{T}} \mathrm{E}_s \left\{ \int_0^{\mathcal{T}} \sum_{n=1}^{N} U\left(x_n\left(s,t\right)\right) dt \right\}. \tag{2.3}$$

The solution to Eq. (2.3) is called a socially optimal solution. For instance, if the communication channels and available resources are static, then Eq. (2.3) can be

written as

$$\max_{s\in\mathcal{S}} \sum_{n=1}^{N} U\left(x_n\left(s\right)\right)$$

without time t.

A solution that maximizes the sum-throughput utility of all UEs might not be practicable, as some UEs might consider it *unfair* in the sense that all resources would be allocated to the UE in the best channel situations. In many environments, *fairness* might be more important than finding solutions providing the largest gains. Various definitions of fair allocations have been proposed, such as weighted fair [52], Max-Min Fair (MMF) [53, 54], and Proportional Fair (PF) allocations [43, 49]. The corresponding utility functions can be written as

$$U(x_n) = \begin{cases} \dfrac{\omega_n}{1-\alpha} x_n^{1-\alpha} & \text{weighted } \alpha\text{-fairness,} \quad (2.4a) \\ x_n & \text{max,} \quad (2.4b) \\ \lim\limits_{\alpha\to\infty} \dfrac{1}{1-\alpha} x_n^{1-\alpha} & \text{MMF,} \quad (2.4c) \\ \log\left(x_n\right) & \text{PF.} \quad (2.4d) \end{cases}$$

The weighted α-fair allocations are a parameterized family of fairness criteria. They are defined for any $\alpha > 0$ and fairness weights $\omega_n > 0$. If $\omega_n = 1$ and $\alpha \to 0$ in Eq. (2.4a), α-fair becomes max-throughput optimization. The max fairness criterion is greedy in that it maximizes network throughput. The disadvantage of such an allocation is that UEs with poor channel conditions are starved of resources. It would seem fairer for all UEs to have access to the network resources. If max-throughput is unfair, then MMF is fairest. In MMF, among all rate allocations, the minimum rate allocated to any flow is maximized over all possible rate allocations, eventually leading to equal rates for all UEs. If $\omega_n = 1$ and $\alpha \to \infty$ in Eq. (2.4a), weighted α-fair allocation reduces to MMF allocation. PF is a compromise between max and MMF. It is based upon maintaining a balance between two competing interests, trying to maximize network throughput while at the same time allowing all UEs at least a minimal level of service. In Eq. (2.4a), the weighted α-fair objective is not defined for $\alpha = 1$, but $\lim_{\alpha\to 1}$ reduces to PF allocation. It is important to remark that if fairness is socially desirable, it is integrated into the utility function. Then social optimality is the outcome in which the utility (sum of players' payoffs) is maximized, and players' payoffs are precisely distributed according to the fairness criterion defined in the utility function.

A network can be modeled using a single function with a utility-based framework, and network resource allocation problems can be studied in a clean manner. The performance of different allocation schemes can be easily compared, e.g., the deviation from a socially optimal solution can be computed. It also aids in examining the trade-off between social optimality and other performance objectives.

B. Grant-free allocation

To measure uplink reliability performance for URLLC, we have considered a grant-free (contention-based) model. Specifically, an uplink transmission scenario is considered where a UE transmits instantly on a time-frequency resource grid whenever data arrives in its buffer, without employing any channel sensing mechanism. The transmission scenario mimics slotted ALOHA [55, 56]. This may result in collisions between packets. System reliability is accordingly measured in terms of the number of transmissions failed due to collisions. For each UE, packet arrival follows the Poisson process with exponentially distributed inter-arrival time [57], with the average number of random access events per access slot denoted by ψ. System BW is divided into K Resource Blocks (RBs) in an access slot. When competing with $N-1$ similar UEs, the UE's collision probability is derived as [PVII]

$$P^{\mathrm{col}} = 1 - \left(\frac{e^{-\psi} + K - 1}{K} \right)^{N-1}. \tag{2.5}$$

For the sake of simplicity, only MAC layer reliability against collisions is considered, and transmission channel impact is ignored in the study.

The contention-based model utilizes various optimization schemes s, which are discussed in Chapter [PVII]. The network optimizes the selected allocation method s where resource utilization is minimized for the given reliability target P^{rel} and latency target T. The number of resources used by the system is then accordingly

$$\begin{aligned} K \quad = \quad & \arg\max_{K} \mathbb{P}_s^{\mathrm{col}}(K, \ell), \\ & \text{s.t. } \mathbb{P}_s^{\mathrm{col}} \leq 1 - \mathbb{P}^{\mathrm{rel}}, \\ & \ell \leq T \end{aligned}$$

where l is the total transmission latency including possible repetitions.

3. Spectrum Allocation

In state-of-the-art mobile communication, exclusive spectrum access is the mainstream approach that national regulatory authorities use to allocate a new spectrum. In exclusive access, only one operator has the right to use a dedicated licensed frequency band. This is undoubtedly needed in 5G mobile systems to guarantee QoS in wide-area RANs, but in scenarios involving small cells deployment with exclusive access, this may result in low spectrum utilization. Non-exclusive access, e.g., in the license-exempt spectrum offers another type of access where multiple operators coexist and utilize the spectrum governed by certain principles related to power restrictions and LBT requirements. Hence, license-exempt solutions can boost the capacity but QoS is unpredictable. Enabling high capacity, guaranteed QoS, and flexible usage is envisioned for 5G systems and thus calls for more flexible regulatory regimes. For instance, operators may utilize various frequency bands with different authorization modes [58] and achieve higher operational BW efficiency and meet the increasing traffic demand on time [59].

Co-Primary Shared Access is a complementary new alternative for spectrum sharing, where multiple operators jointly use a part (or the whole) of their licensed spectrum [58, 60]. One of the most relevant Co-Primary Shared Access scenarios is the Limited Spectrum Pool (LSP). In the LSP scenario, a limited number of operators share a common or mutual resource pool by relying on more flexible and adaptive prioritization policies than is currently possible with dedicated licenses [61]. Cognitive radio technologies are examples of effective policy measures to resolve sharing conflicts over the LSP under vertical spectrum sharing [62], where the lessor (owner) operator has higher legacy rights over the spectrum than the lessee operator. On the other hand, the co-primary or horizontal spectrum sharing scheme conceptualizes the case where authorized operators possess equal ownership on the spectrum being adopted [60]. Investment decisions under co-primary shared access have a lower risk, if the set of peer operators and principles of spectrum usage are known beforehand.

The multilateral use of shared resources in the LSP can, for instance, be achieved

with channel allocation schemes developed initially for single-operator systems. These schemes are in principle applicable to realize inter-operator spectrum sharing, provided that operators are willing to exchange information and cooperate honestly. With such stipulation, literature discusses various sharing algorithms extensively. However, they differ in the domain where inter-operator interference is handled, i.e., time, frequency, or space.

Joint use of licensed spectrum among operators can be realized either orthogonally in time [63], frequency [64, 65], space, or non-orthogonally [65, 66]. In cooperative orthogonal time domain sharing, operators with a low load can borrow their time resource to heavily loaded operators, helping them to reduce blocking probability and frame delay [63]. An upper bound for sum capacity in a scenario with two operators sharing frequency resources orthogonally is presented in [64], where operators have full access to UE-specific channel quality indicators of all shared channels to perform coordinated scheduling. In [65], orthogonal sharing based on a pairwise exchange of RBs between two operators is considered.

In non-orthogonal spectrum sharing, operators simultaneously use the same spectral resources, creating inter-operator interference. In [66], a cooperative game is proposed, which requires full knowledge of action profiles. Non-orthogonal inter-operator spectrum sharing in a spatial domain has been considered in [65], where CSI is exchanged among operators to implement coordinated transmit beamforming and steer their antenna beams towards the most convenient direction. In all these studies, operators benefit from cooperative spectrum sharing; however, operators need to reveal proprietary information to their competitors [65, 66] or central entity [63, 64]. Operators are competitors in nature, and the rationale to cooperate would be either a legal framework or self-interest. There is no reason to assume that operators are willing to exchange proprietary information with their competitors in such settings. Also, if competitive operators choose to exchange information, they may have an incentive to misreport it. Besides, information exchange may also carry excessive inter-operator signaling overhead. Operators provide differentiated services to their customers, and the objectives can be categorically different according to their business models. Thus, optimization of a joint utility by a central entity is not realistic. Co-primary spectrum sharing between operators is thus characterized by operators not knowing each others' optimization targets or network states.

Spectrum sharing can also be realized by monetizing spectrum usage and arranging auctions to determine spectrum management [53]. In [48], spectrum broker is considered to chair the bidding. Here we are interested in adaptive spectrum use on a short time scale, e.g., related to cell load variations and changing inter-operator interference conditions. It is a non-trivial task to design efficient auction mechanisms

for a limited area and time and to couple operator's auction strategies to their income model. Further, implementing auctions also requires the involvement of a trusted third party with substantial accounting infrastructure to collect bids and track payments. Besides, operators may be hesitant in adopting market-driven sharing schemes as they may not want to touch their revenue model. Accordingly, we shall consider non-monetized spectrum sharing, happening directly between RANs of operators.

The interaction between self-interested operators can be modeled by non-cooperative games, where operators make decisions independently. A non-cooperative one-shot game formulation for licensed and unlicensed (license-exempt) access is discussed in [67] and [68], respectively. In [67], operators enlist their preferences of splitting up the shared pool, and the outcome is established based on a minimum rule. This method may not work well in scenarios with load variations, in which heavily-loaded and lightly-loaded operators will end up with the same number of orthogonal carriers. Whereas in [68], operators are not constrained by any rules and freely select power allocation strategies during repeated games. The parameters such as power constraints and channel gains are common knowledge to all the system. In multi-operator spectrum sharing, utilities and strategies are operators-specific and are not shared among them. There may also be specific rules, either agreed among operators or enforced by a legal entity. Operators' strategic choices are bounded by these rules where no deviations from the rules are allowed. For instance, the legal framework governing operation of different systems in license-exempt bands (e.g., Wi-Fi and Zigbee) gives equal access rights to all radio devices when complying with certain power emission levels, spectral masks, channel reservation protocols, and activity rates in case of Industrial, Scientific, and Medical (ISM) bands.

Considering long-term sharing in license-exempt bands, repeated games strategies that lead to favorable Nash Equilibrium (NE) are also discussed [68,69]. These games have no a priori rules. It is assumed that nodes exchange information and agree in advance on their operational points regarding power allocation across shared frequency channels, e.g., at a NE in [68] or employing orthogonal allocation in [69]. The agreement is then enforced under the threat of punishment. Obviously, without exchanging information about network states and optimization objectives, it is not possible to identify and punish cheating, and cheating as such becomes ill-defined. In [70], operators establish cooperation and play non-zero-sum games to share spectrum. However, the utility function based on spectrum pricing is undesirable as it punishes the operator for its increased spectrum usage. Thereby, in a co-primary spectrum sharing setting, the applicability of strategies based on concepts of dishonesty and punishment may be limited.

Among competitive operators, a non-cooperative game-theoretic solution appears

to be a more viable option for spectrum sharing. Unlike spectrum sharing in license-exempt bands, where the number of players sharing spectral resources is indefinite, we consider a fixed and known number of players with publicly known and persistent identities. Operators make decisions independently, and the cooperation is entirely self-enforcing. Operators do not share proprietary information, e.g., network states and optimization targets, with each other or external parties, and because of that, it is not possible to agree in advance on the operational point. Also, there is no reason to assume that operators are willing to take part in monetary transactions. Under these constraints, we set up a repeated games model in which operators may take and grant spectrum usage favors at each stage game. The term 'favor' captures the notion that the operator has given a favor to someone else and expects the beneficiary to do the same in the future [71–73]. A favor refers to the case where the operator asks permission to use resources from the spectrum pool exclusively. We assume no other intra-RAN information exchange but limited message exchange among operators to realize spectrum negotiations through favors. It requires a new interface, which may be over-the-air or over the core network. For instance, an operator with a high network load may ask for spectrum usage favors, and another operator with few UEs to serve may vacate some resources from the pool for some time. Unlike spectrum sharing based on a one-shot game, in the repeated games model, the action of operator at each stage game takes into account immediate rewards and history of interactions with the opponent operator entailing benefits of reciprocity.

Next, we describe game theory fundamentals and game-theoretic inter-operator spectrum sharing protocols proposed in [PI, PII, PIII].

3.1 Introduction to game theory

Game theory is the study of mathematical models predicting the outcome of strategic interactions [74]. It analyzes or models interactions between interdependent decision-makers that may have self-/mutual interests and conflicts. Game theory was increasingly recognized in the early 20^{th} century, primarily in economics describing competitor's behavior. The game-theoretic interactions are central to understanding various fields, e.g., political sciences, logic, psychology, and biology. Telecommunications is one of the emerging fields where game theory has gained attention to model conflicts among players in the scenarios involving, such as dynamic spectrum sharing, congestion, trust management, routing, power, and topology control [73]. The importance of modeling interaction using game theory is diverse. It offers a range of optimization criteria, e.g., single or multi-stage games. Players can devise strategies independently, and no central control is needed. Cooperation can be enforced if

players have mutual interests, even though players are autonomous.

Game Definition

A game $\mathcal{G}$ is a triple of sets of players $\mathcal{P}$, strategies $\mathcal{S}$, and utility functions $\mathcal{U}$, i.e.,

$$\mathcal{G} = \langle \mathcal{P}, \mathcal{S}, \mathcal{U} \rangle \ .$$

A finite set of players, $\mathcal{P} = \{1, 2, 3, ..., m\}$, occupies an m-tuple of pure strategy sets, one for each player, i.e., $\mathcal{S} = \{s_1, s_2, s_3, \ldots, s_m\}$, where s_i is strategy profile of i-th player, such that $s_i \in S_i$ and S_i is a finite number of the allowable strategy set of i-th player. Utility function represents player's award at the outcome of game, i.e., $\mathcal{U} = \{U\left(s_i, s_{-i}\right)\} \rightarrow \mathbb{R}$, for the i-th player strategy s_i and other players' strategies represented with s_{-i}. Players may behave greedily if strategies conflict, i.e., utility gain for an individual may result in a loss for the opponent. Hence, an individual's choice of strategies may have short or long-term repercussions. For this reason, the choice of strategies should be rational than pure greed if interactions persist.

Game Strategies

Players can adopt various kinds of strategies which may result in different equilibria or sub-/optimal game resolutions. Some notable strategies are described as following.

1. Dominant strategy: A strategy $s_i^* \in S_i$ is a dominant strategy to a given strategy $s_i^{'} \in S_i$ for player i if $\forall s_{-i} \in S_{-i}$ when

$$U_i\left(s_i^*, s_{-i}\right) > U_i\left(s_i^{'}, s_{-i}\right).$$

In the prisoner's dilemma example depicted in Fig. 3.1, for any opponent $-i$'s action, non-cooperative action brings the biggest payoff for player i. Hence, (non-cooperation, non-cooperation) becomes the dominant strategy.

2. Pareto optimal: In a game, strategy profile s^* Pareto dominates strategy profile s when

$$U_i\left(s_i^*, s_{-i}^*\right) \geq U_i\left(s_i, s_{-i}\right) \quad \forall i \in \mathcal{P},$$

and there exists some $j \in \mathcal{P}$ for which

$$U_j\left(s_j^*, s_{-j}^*\right) > U_i\left(s_i, s_{-i}\right).$$

This strategy profile s^* is Pareto optimal only if no other strategy profile s' Pareto dominates s^*. In other words, a strategy profile s^* is said to be Pareto optimal if we cannot find another strategy profile s in which it is impossible to make any player better off without making at least one player worse off. In the prisoner's dilemma example, except dominant strategy, i.e., (non-cooperation, non-cooperation), all the other three strategies are Pareto strategies. The reason is that when a player jumps from a Pareto strategy to any one of the other three strategies, the payoff of at least one

player gets worse. Whereas for (non-cooperation, non-cooperation) strategy, both players can improve their payoffs by switching to (cooperation, cooperation) without making anyone worse, and thus (non-cooperation, non-cooperation) is not a Pareto optimal strategy as it is Pareto dominated by (cooperation, cooperation) strategy.

3. Best response: A strategy $s_i^* \in S_i$ is a best response for player i to $s_{-i} \in S_{-i}$ if $\forall s_i \in S_i$,

$$U_i\left(s_i^*, s_{-i}\right) \geq U_i\left(s_i, s_{-i}\right).$$

The best response is different from the dominant strategy in a way that the best response improves utility for a specific strategy $s_{-i} \in S_{-i}$ and $\forall s_i \in S_i$, whereas the dominant strategy improves utility to a given strategy $s_i^{'} \in S_i$ and $\forall s_{-i} \in S_{-i}$.

4. Nash equilibrium A strategy profile s^* is a NE, if $\forall i \in \mathcal{P}$, s_i^* is the best response strategy to s_{-i}, i.e., no player can do better unilaterally changing its strategy. In other words, a dominant strategy equilibrium is always a NE, which is (non-cooperation, non-cooperation) strategy in the prisoner's dilemma example.

Game classifications

Players' strategies are highly influenced by regulations imposed by the nature of the game's environment. It can drive game model taxonomies in several ways. Some notable classifications are discussed as follows.

1. One-shot vs. repeated games: One-shot or single-stage game is played only once. Players may be uninformed about others' moves, and they might act selfishly to attain the highest payoff. Risks are high but carry no further repercussions. On the other hand, repeated games are played numerous times where each player's strategy to be contingent on past moves and has reputation effects and retribution for it. Further, repeated games can be of finite and infinite occurrence. In infinite repeated games, according to the Folk theorem, there exists a discount factor $\widehat{\delta} < 1$ such that any feasible and individually rational payoff can arise as an equilibrium payoff for any discount factor $\delta \in \left(\widehat{\delta}, 1\right)$. Thus, future payoffs are discounted and are less valuable. It is because consumption in the future is considered less valuable than present due to time preference (e.g., money). The player's total payoff in repeated games is a discounted sum of each stage payoff. The repeated games model holds a variety of equilibrium properties because the threat of retaliation is real due to the repetitive nature of the game, and also, it has a much larger strategy space than the one-shot game. Unlike the one-shot game, players can punish hostile players using tit-for-tat strategy in repeated games [75]. Repeated games with perfect monitoring where players' actions are observable is called a multi-stage game. In this, players announce their strategy publicly and thus each stage of a multi-stage game resembles a single-stage game.

2. Cooperative vs. non-cooperative game: A cooperative game is defined as where a group of players enforces cooperative behavior on each other. In this, players bargain or negotiate on payoffs and form joint strategies. A cooperative game is pragmatically undesirable because of excessive overhead signaling and trust issues, but it provides a unique Pareto optimal solution for problem modeling. In the latter, if competition is between potentially conflicting and self-interested players, then the corresponding game is known as a non-cooperative game. In a non-cooperative game, without centralized control, players do not cooperate or make deals where any cooperation among them must be self-enforcing motivated by their self-interests.

3. Zero- vs. non-zero-sum game: A zero-sum game is a game played between multiple players in which the total gains by the players are precisely the losses borne by the rest of the players, e.g., in gambling. In non-zero-sum games, the sum of utility gains and losses of winning and losing players is not zero, which is typically witnessed in spectrum sharing games.

Prisoner's dilemma example

The prisoner's dilemma is a paradox in decision-making which describes a situation where two prisoners are taken into custody in connection with a burglary. The authorities possess insufficient evidence to convict them for their crimes and rely on statements from prisoners, which can influence their sentences. Nicknamed in 1950 by Albert W. Tucker, prisoner's dilemma is a standard example in game theory. Fig. 3.1 shows the prisoners' strategies, i.e., confess or lie and their outcomes in terms of sentences.

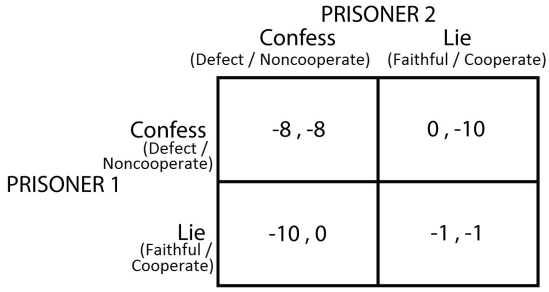

Figure 3.1. Prisoner's dilemma game.

If both prisoners lie and help each other, they will be charged with the lesser term of a year. The authorities question them separately, which means that prisoners make decisions unaware of the other prisoner's choice. The process is non-cooperative with imperfect information. The authorities will try to convince each prisoner to confess a crime by offering an escape clause and to accomplice a prison term of ten years. If both prisoners defect and confess, they shall be sentenced to eight years.

Both prisoners can assume that both have been offered the same deal but are entirely unaware of each others' choices. In this, (negative of) prison term can be seen as a utility for each choice, where each would like to have maximum utility, i.e., a minimum prison term. With no repercussions, prisoners are tempted to get away with a minimum prison term, and therefore will defect. Consequently, "to confess" is the dominant strategy. Thus (non-cooperation, non-cooperation) is a NE in this one-shot game.

Let us assume two burglars work together over a long time and repeatedly get interrogated and sentenced. After each round of interrogation and subsequent sentencing, they get to know each other. As the game is no longer one-shot, prisoners would like to have minimal prison terms for their repeated criminal acts. So, rationally thinking, it makes sense to cooperate over time, unlike they behave in a one-shot game, and both prisoners can get away with shorter sentencing of a year term. Let us say prisoner $P1$ wants to defect, and no doubt he can get away with a zero prison term, whereas prisoner $P2$ will be awarded ten years of prison term. However, prisoner $P2$ will become aware of the last hostile choice made by prisoner $P1$, and therefore will defect too in the future to punish prisoner $P1$ for the non-cooperation and non-trusting behavior. If they both continue with their non-cooperative behavior, they will always get eight years of prison term for every subsequent crime. It shows that prisoner $P1$ may have a temporary benefit with no sentence, but in the long run, $P1$ may lose by having a much longer sentence of eight years every time for repeated criminal acts. Therefore, in repeated games, "to lie" is the dominant strategy, and (cooperation, cooperation) is a NE where rational player tends not to deviate due to fear of punishment with tit-for-tat strategy from the opponent [75].

3.2 One-shot game model

Two operators, $X_{\{A,B\}}$, are interested in sharing spectrum and thus agree to play one-shot or repeated games. Before starting with either game, operators first agree on default allocation, and each operator divides its BW into $K + K_D$ CCs. Under the LSP setting [58], they agree to share K CCs and inflict interference on each other due to shared use, and while on K_D CCs, they transmit exclusively. Given load, BW, channel quality, and inflicted interference on shared CCs, each operator measures its PF Utility U as per Eq. (2.1), (2.4d) described in Chapter 2.

We first consider a one-shot game presented in [PI] where two myopic operators play the non-cooperative game and are interested in increasing their utility. In the game, they do not consider their past actions while choosing their current actions. The action or strategy s_X of each operator considers either asking a favor of $k_a =$

$1, \ldots, K$ CCs or granting a favor on $k_g = 1, \ldots, K$ CCs. To specify the game's outcome, when a player asks a favor on k_a CCs, the other has granting ability on k_g CCs, and if $k_a \geq k_g$, then the outcome is an exchange of k_g CCs where the granting player stops transmitting on k_g CCs.

In this game, an outcome is a NE where no player can improve its utility by deviating unilaterally, i.e., $U_X\left(s_X, s_{-X}\right) \geq U_X\left(s_X^*, s_{-X}\right)$ for every alternative strategy s_X^*. It means a NE of the one-shot game corresponds to the situation where both players always ask a favor on K CCs to maximize its reward but never grant. Consequently, both operators would utilize all K CCs from the LSP irrespective of network load and inflict interference continually.

A similar game is formulated in [68] to water-fill the available power on the unlicensed (license-exempt) spectrum using a one-shot game. While playing strategies, [68] assume the understanding of the opponent's utility value. However, the games discussed in this thesis (one-shot or even repeated) consider no revelation of utility or related information, and operators are free to construct any utility function. Nonetheless, the result would be the same, i.e., fully utilizing the available spectrum, which is the only option for a rational player under the myopic setting.

3.3 Repeated games model

The operators are interested in sharing spectrum for a long time and thus, they will be focused on playing the above described one-shot game repeatedly. A player's action at a stage game depends not only on the current actions but also on the sequence of previous actions where an action amounts to gain or loss in utility. It admits a rich set of strategies and equilibrium profiles, including punishment strategies, which is well demonstrated with the Prisoner's dilemma classic example [74]. The repeated games with a focus in this thesis are even more challenging as the game is a *stochastic game* meaning that each player's payoff depends on a random parameter, which influences load and channel characteristics. Further, players have *imperfect information* since they do not reveal critical information to each other and thus the game can be observed as a *Bayesian game*.

Operators play repeated games [PI], where at each stage of the game, each operator computes its utility gain and loss by asking and granting favors on $k = 1, \ldots, K$ CCs. The operator draws Probability Density Functions (PDFs) of utility gains f_{X,G_k} and losses f_{X,L_k} over k CCs. During each stage, the operator first checks whether to ask for the biggest favor, i.e., a favor on K CCs, by comparing its utility gain with the corresponding threshold $\theta_{X,K}$. If utility gain is relatively lower, the operator checks for a favor on $(K-1)$ CCs instead and compares with threshold $\theta_{X,K-1}$, and so

forth. As a result, the probability of asking favor on k CC is derived as

$$P_{X,k}^{\text{ask}} = \prod_{j=k+1}^{K} \int_0^{\theta_{X,j}} f_{X,G_j}(g_j)\, dg_j \int_{\theta_{X,k}}^{\infty} f_{X,G_k}(g_k)\, dg_k, \tag{3.1}$$

given that favor cannot be asked on $k+1, \ldots, K$ CCs. Similarly, an operator grants a favor on k CCs if its immediate utility loss is smaller than threshold $\lambda_{X,k}$, and it has not requested a favor on any of k CCs. Thus, the probability of granting a favor on k CCs is

$$P_{X,k}^{\text{grant}} = \prod_{j=1}^{K} \int_0^{\theta_{X,j}} f_{X,G_j}(g_j)\, dg_j \int_0^{\lambda_{X,k}} f_{X,L_k}(l_k)\, dl_k. \tag{3.2}$$

We assume similar operators, and therefore, they will exchange an approximately equal amount of favors averaged over a long period. It is inspired by [76], which provides a preliminary understanding of steady-state behavior in this type of setting. Favors would only become a virtual RAN-level spectrum sharing currency. Thus, a sharing constraint is defined as

$$\sum_{k=1}^{K} k P_{A,k}^{\text{ask}} P_{B,k}^{\text{grant}} = \sum_{k=1}^{K} k P_{B,k}^{\text{ask}} P_{A,k}^{\text{grant}}, \tag{3.3}$$

where the left-hand side and the right-hand side describe the average number of CCs that Operator A and B receive their favors on, respectively. The average received favors would be similar for similar operators. However, it is perfectly okay to consider dissimilar operators. Then the favors request would be dominated by a demanding player. It may result in a favors exchange gap at the end of repeated games. It means a relatively less demanding, rational player will agree to play games only if it foresees a profit. For example, this can happen if this less demanding player expects to receive favors during extreme urgent times, and the player should be compensated additionally with money or some form of settlement for skewed usage.

The operator monitors asking and granting probabilities of the opponent and identifies its decision thresholds for satisfying the constraint (3.3), which maximizes an excess utility $\widetilde{U}$. The excess utility for an operator is its expected gain from taking favors penalized by its expected loss from granting favors, which is

$$\begin{aligned} \widetilde{U}_A = \sum_{k=1}^{K} \Bigg(& P_{B,k}^{\text{grant}} \prod_{j=k+1}^{K} \int_0^{\theta_{A,j}} f_{A,G_j}(g_j)\, dg_j \int_{\theta_{A,k}}^{\infty} g_k f_{A,G_k}(g_k) dg_k \\ & - P_{B,k}^{\text{ask}} \int_0^{\lambda_{A,k}} l_k f_{A,L_k}(l_k)\, dl_k \prod_{j=1}^{K} \int_0^{\theta_{A,j}} f_{A,G_j}(g_j)\, dg_j \Bigg). \end{aligned} \tag{3.4}$$

In order to solve the optimization problem that maximizes excess utility, the Lagrangian function is constructed and solved for first-order conditions. The mathematical procedure is well described in [PI] in obtaining the thresholds $\lambda_{A,k}, \theta_{A,k} \forall k$

that maximizes the Lagrangian and satisfies the constraint (3.3). The obtained system of equations does not accept a closed-form solution and is solved numerically. Further, in the Appendix in [PI], it is shown that excess utility is always positive for the obtained set of thresholds. It is indicative of cooperation in this repeated games setting for inter-operator sharing is a NE. A similar approach is described in [77], a particular case with the negotiations conducted over $K = 1$ CC.

Next, the performance of the repeated games model is assessed in an indoor deployment scenario and depicted in Fig. 3.2. The modeling parameters are detailed in [PI]. First, an initialization phase of 200 000 simulation snapshots (or 200 000 stage games) is considered in the sharing game. At each stage game, UEs' locations are independently generated according to the PPP by varying its mean. In this manner, the distributions of utility gains and utility losses are tracked and captured at the end of the initialization, see Fig. 3.3 for the LSP with $K = 2$ CCs which can be seen as the steady-state distributions over all possible network states.

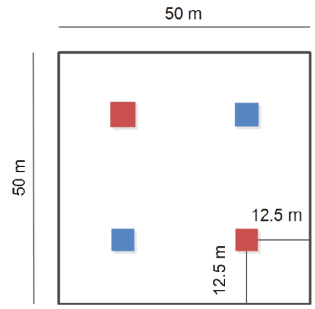

Figure 3.2. Indoor inter-operator deployment scenario, where different colors represent gNBs of different operators [PI]. © [2015] IEEE

The gains due to the exchange of favors are demonstrated with operators having an equal mean network load of five UEs. The repeated games model takes advantage of instantaneous network load variations. The low load operator shuts its transmission on the desired number of CCs upon being asked by the high load operator during a stage game, with an intent of expectation of reciprocal behavior from the high load operator when load situations reverse in the other stage game. The gains are compared to NE of a one-shot game where both operators always utilize the whole LSP and inflict interference. Each operator improves its mean UE rate by approximately 25 %, depicted in Fig. 3.4 over LSP with $K = 4$ CCs. Further, in [PI], rate gains are depicted in scenarios with asymmetric mean load and over LSP with lesser CCs.

The repeated games model is also replicated in a Heterogeneous Network (HetNet) scenario [PII], shown in Fig. 3.5. Each operator serves a geographically separated building with multiple femto gNBs. It can be motivated by the fact that these different buildings have most UEs belonging to a single operator. However, there will be a

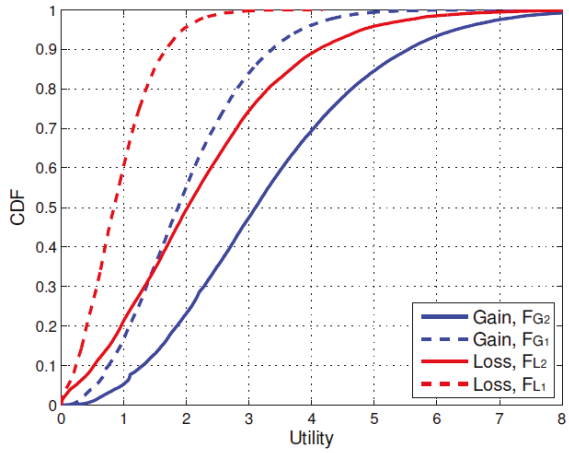

Figure 3.3. Distribution of utility gains and losses for an operator at the end of the initialization over LSP with $K = 2$ CCs [PI]. © [2015] IEEE

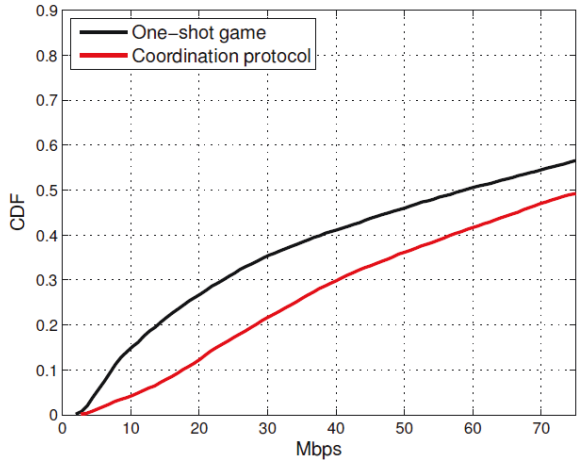

Figure 3.4. UEs' rate distribution of an operator over LSP with $K = 4$ CCs and equal mean load for operators [PI]. © [2015] IEEE

minority of UEs in each building with mobile connections with the operator serving the other building. Therefore, the operator deploys a micro gNB outside its serving building to provide services to its minority of UEs in the other building. Operators allocate an equal amount of BW for their micro and femto nodes, where each operator's own micro and femto UEs do not interfere due to geographical separation and micro gNB's antennas orientations.

Each operator's majority UEs are expected to be served with femtocells. Thus, operators might be interested in utilizing other operator's BW resources for their femto-connected UEs. The UE rates can be improved with the doubling of femtocells' BW resources compared to no sharing. Due to sufficient separation between femtocells of different operators, the interference between femto-connected UEs of operators is

decoupled and negligible. Thus, operators benefit from doubling resources without any notable increase in interference from the other femtocells, and rates may become almost double. Hence, operators agree to share their BW resources and form an LSP, and the default resource arrangement is shown in Fig. 3.5.

The only danger is that when the operator's minority UEs are visiting the other building, its micro-connected UEs can interfere with femto-connected UEs of the other operator. In such situations, if interference is excessive, the operator serving micro-connected UEs would be interested that the other shuts its femtocells' transmission on some of shared CCs. With this, micro-connected UEs will be served in an interference-free environment, whereas femto-connected UEs of the other operator will already have at least half of the BW resources from the LSP. Hence, operators will be interested in playing repeated games, where an operator can ask spectrum favors on $k = 1, \ldots, K$ CCs if utility gain is more than the gain threshold, and the other operator grants if its utility loss is under the loss threshold. It is plausible that the operator's request for a favor can get rejected, and micro-connected UEs suffer heavily because femtocells of the other operator remain active in shared BW. Therefore, to provide a safety net to micro-connected UEs, at least one CC is allocated exclusively; see resource arrangement depiction in Fig. 3.5. The game behavior is similar to the one which is presented before in the non-HetNet setting. The primary difference is utility calculation in the HetNet, which is the sum of microcell and femtocells utilities. The procedure for derivation threshold values would be the same and is covered in detail in [PII].

The performance in the HetNet setting will depend on the UE's visiting probabilities in the other building. If the operator's visiting UEs are set to become large, then the operator would perhaps may not even agree to play the game, as both of them will want each other to shut their shared CCs used by femtocells. In these scenarios, operators should keep their spectrum orthogonal and never form an LSP. The game is interesting if visiting probability is low, and the rate performance is depicted in Fig. 3.6 for a scenario with a low visiting probability of around 20 %. It clearly shows that femto-connected UEs benefit from extra BW resources, and the mean UE rate is improved by almost 80 %. The low rate UEs, which are probably mico-connected UEs, do not degrade substantially. During the pressing times, operators can secure favors and limit interference partially. UE rates with high visiting probability are also depicted in [PIV]. As described earlier, gains are non-existent in such settings.

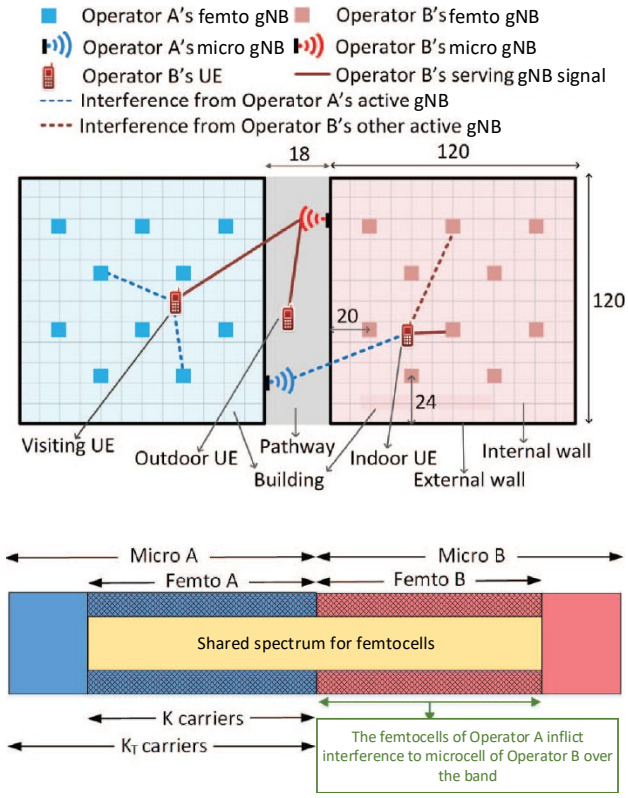

Figure 3.5. Indoor inter-operator HetNet and default LSP scenario [PII].

3.4 Coordination protocol

In this section, the coordination protocol presented in [PIII] is discussed, which tends to exploit both short-term and long-term reciprocity. Through coordination protocol, two peer operators handle spectrum negotiations. Unlike in Section 3.3, where default resource allocation is LSP, the coordination protocol begins with a default allocation of Mutual Renting (MR) [58], depicted in Fig. 3.7. It is a more reasonable assumption as it depicts the actual resource allocation stance where two operators do not share spectrum.

In this coordination protocol, operators play both one-shot and repeated games to settle on with spectrum favors negotiations. The operators' first attempt is to play a one-shot game, where one-shot game's strategies are depicted in Fig. 3.7. The one-shot game protocol is inspired by [67], where operators tend to agree on symmetrical resource allocation, i.e., both will utilize an equal amount of BW resources. Once the one-shot game outcome is reached, operators play repeated games, and the relative demanding operator tries to secure more spectrum favors. Fig. 3.7 shows both one-shot and repeated games' actions in the MR scenario, along with all probable

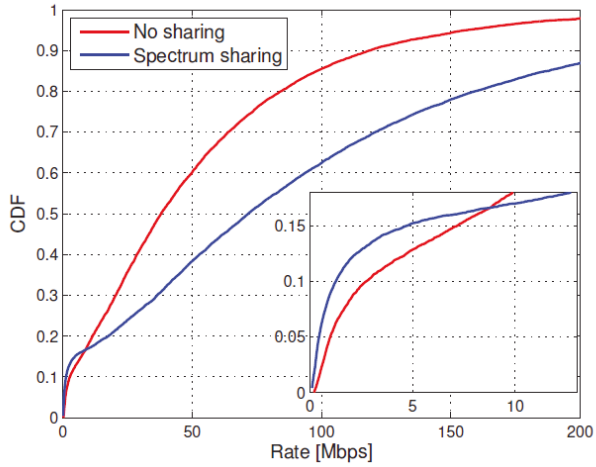

Figure 3.6. Rate distribution of an operator with UE's low visiting probability in HetNet [PII]. © [2015] IEEE

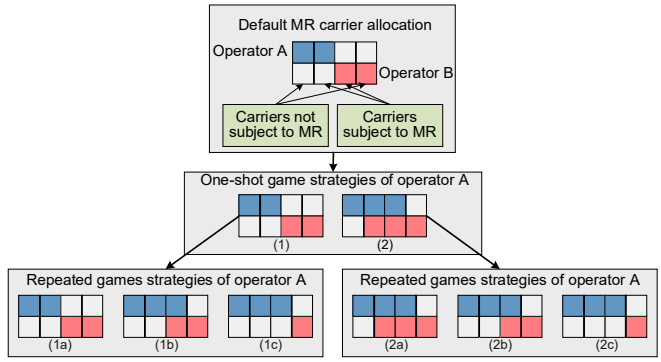

Figure 3.7. One-shot and repeated games strategies in MR scenario [PIII].

allocations in a stage game. Fig. 3.8 shows the performance curve when operators share spectrum through the coordination protocol. The simulation considers unequal load settings in high interference scenarios. The one-shot game only benefits when UEs are not facing interference, and as a result, operators tend to increase common resource usage, and gains are reflected among high rate UEs in the figure. On the other hand, repeated games act when interference is high and loads are uneven. The more demanding operator can further secure more favors, and it improves rates of inferior UEs witnessed in the bottom tail of CDF. Hence, the total gains from one-shot plus repeated games surpass one-shot game alone but fall behind the cooperative game, which is expected. Both 10-th and 50-th percentile UEs improve their rates by approximately 50 %. In [PIII], results for equal mean load under low interference conditions are also presented. The one-shot game enables the operators to share the whole spectrum (as interference is low), and this is the same achieved with cooper-

ative sharing. The repeated games have no impact as the upper bound of gains is already reached with a one-shot game alone.

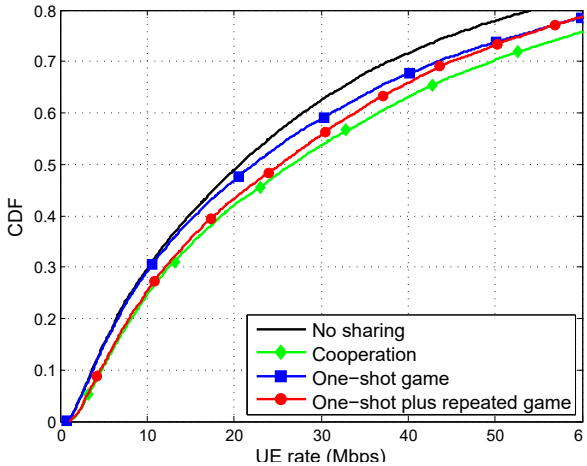

Figure 3.8. UE rate distribution in high interference and unequal mean load scenario.

3.5 Summary

In this chapter, we consider the resource allocation problem to deliver efficient broadband services. The problem is examined in a framework consisting of two similar but competitive operators interested in sharing spectrum. The sharing problem is modeled as a non-cooperative two-player game, which is played by exchanging spectrum favors when one of them has relatively more substantial needs. The game requires no revelation of operators' critical information to each other, or to any third-party entity. Each operator sets two thresholds, determining the conditions for borrowing and granting favors. Operators choose their actions independently. They are free to act in their self-interest, e.g., they can deny opponents' requests or always demand favors. The game will thrive when both operators behave rationally and reciprocate favors, especially in situations or stage games with imbalance loads. The game is repeated over many stages constrained by the net exchange of favors amounting to approximately zero at the end of repeated games. Operators focus on maximizing net gains from getting and giving favors, and the game converges to Nash equilibrium. The game provides gain relative to a one-shot game or no sharing in both indoor and heterogeneous deployment scenarios. Gains are further improved if strategies are based on combinations of one-shot and repeated games. The cooperative game, requiring the revelation of critical information related to traffic or load, acts as an upper bound. It is not a preferred choice for competitive operators which lack trust in each other.

4. Ultra-Reliable, Low Latency Resource Allocation

5G is envisioned with new services and use cases are envisioned [2]. These new services are not only for human communication; the significance of MTC is predicted to grow. It brings significant improvement to communication between machines. Initially, the goals of 5G research concerning machine communications were to enable high reliability, low latency communications, and a possibility to connect a vast number of devices with low energy consumption [59]. Latter has been realized to a great extent already as part of 4G LTE via the Narrowband-Internet of Things (NB-IoT) technology, and further enhancements are slated in future releases of 5G NR. URLLC is one such usage scenario that enables mission-critical MTC, including real-time control and automation of dynamic processes in, e.g., factory automation, smart grid protection, traffic management, and safety or "Tactile Internet" [2, 78, 79]. Some aspects of URLLC are included already in the initial 5G NR releases of the 3GPP. Nonetheless, significant additional work remains in further stages of the evolution of 5G.

Earlier generations have steadily improved latency, achieving 10s of ms in LTE well suited to serve the needs of human-to-human communication. However, in many use cases, machines can significantly benefit from reliable, lower latency communications. The most stringent reliability requirement on URLLC in NR Release 15 is 99.999 % under radio latency bound of 1 ms [2]. The maximum Packet Error Rate (PER) must not be higher than 10^{-5}, where maximum allowable latency, including jitter/retransmissions, is down to 1 ms. LTE also offers Guaranteed Bit Rate (GBR) that can support up to 10^{-6} PER, applying Radio Link Control (RLC) Automatic Repeat Request (ARQ) and HARQ protocols. The delay budget, however, goes until 300 ms [80]. This delay includes transport and core network latencies. In RAN, this reliability is achieved with ARQ and HARQ retransmissions, each round-trip taking at least 8 ms. Thus, LTE systems may guarantee 99.999 % reliability, but the delay budget far exceeds 1 ms. These solutions are accordingly unfit for URLLC. However, with new numerology consideration for 5G NR [16], e.g., reduced slot

size, non-slot-based scheduling, and reduced processing time, there is a possibility to support reliable transmissions with 1 ms latency.

There are three aspects to reliable use of communication; latency, reliability, and availability [82]. It is essential to distinguish between reliability and availability. Reliability is related to the short-term quality of communication and is affected by multipath fading and rapidly varying interference. Availability characterizes service enjoyed by an end-UE in the longer term. It is affected by path loss, shadowing, and other slowly varying characteristics of the channel. Intuitively, service availability is a geographic characteristic, which varies from place to place and potentially on long time scales, e.g., related to hardware failures. Reliability of service is a temporal characteristic related to varying service quality and delay. In the literature on mission-critical MTC, concepts related to availability and reliability are often not distinguished. They are referred to with the common term "reliability".

To provide highly available wireless connectivity, we need an increased immunity to interference and environmental clutter besides the densification of nodes. Rapid fluctuations in a channel and inter-cell interference may be countered using MIMO technologies. State-of-the-art methods to ensure reliability are based on diversity. Conventional methods to create temporal, spatial, and multipath/frequency diversity work against randomness in fast fading but do not help against the outage caused by shadow fading. To combat shadow fading, routing diversity may be exploited by having multiple alternate communication routes to choose from. A few alternatives can be used for this. Macro diversity, where a UE is simultaneously connected to multiple gNBs, is one alternative.

Also, multi-connectivity [83] is a potential solution, where a UE has multiple connections on different carriers, possibly to different gNBs. Different Radio Access Technologies (RATs) are used on carriers, and they have different propagation characteristics. In [84], reliability against fast fading using spatial diversity and HARQ retransmissions was addressed. In [85], robust link adaptation for URLLC for mobile terminals was considered. Reliability improvement strategies influenced by BW requirement, antenna configuration, latency budget, interference power level, channel estimation, diversity order, and deployment scenarios are discussed in [86–88]. Additionally, path loss and shadow fading pose severe problems as well. In [89], multi-hop communication for creating routing diversity was considered for URLLC. A single RAT is used, and a single gNB controls communication with the UE, choosing from a set of potential Relay Nodes (RNs) to help with communication. Tentative results with 99.9 % reliability and 95 % availability were presented. A system with a single UE per cell was considered so that inter-cell interference was static.

There has also been renewed interest in contention-based access in 5G systems

to ensure low latency access, which was initially proposed for mMTC. Transmission resource is organized in contention periods, where UEs compete and access the resource. It requires less signaling than allocation methods involving SRs, and therefore, latency is saved. On the other hand, this allows collisions among UEs if they access the same resource. Hence, it may be preferable for sporadic small payload data to send the data packet at the very first step, alongside minimal control information. In a typical SPS allocation, the UE can transmit data readily without sending SR every time, e.g., in voice communication. However, in SPS, a dedicated resource is usually allocated for the UE to transmit. Therefore, it can lead to enormous resource wastage if traffic is random and sporadic, and contention-based access, perhaps, would be more spectrum efficient.

Contention-based schemes can be classified as single and multi-channel slotted ALOHA protocols [55, 56, 90–93]. For example, the LTE or NR Random Access Channel implements multi-channel slotted ALOHA, where UEs compete for access slots in a time-frequency resource grid for setting up connections. In our work, we are interested in similar approaches for data transmissions that are subject to extremely high reliability requirements, e.g., 99.999 %, under a tight latency window for URLLC. In [90], multi-channel slotted ALOHA is analyzed for a fixed spectrum for multi-channel satellite communication. In [91], UEs access the channel with some probability, which is broadcast by the network to prevent collisions. However, discussed schemes do not follow tight latency requirements. Moreover, [92, 93] considered successive interference cancellation to retrieve collided packets, thus increasing the latency.

In [94, 95], contention-based schemes are enhanced by allocating a combinatorial pattern to transmissions. Eq. 3 in [94] shows that allocating a deterministic pattern to a UE's transmission repetitions instead of a random pattern will result in a lower collision rate. However, the patterns are allocated in the time domain, resulting in some UEs having longer incurred latency. A similar approach can be utilized to allocate frequency hopping pattern for repetitions in consecutive slots. The drawback is that whenever contention resource size is changed, all UEs will need to be informed of their updated hopping patterns, causing multi-fold utilization of DCI resources. A similar scheme is discussed in [95], where the UE transmits two transmissions associated with a HARQ process with a specific pattern. If both transmissions collide, the gNB can identify the UE with a viable probability due to its transmissions' pattern. Nonetheless, it suffers from the same drawback as [94].

In this thesis, we inspect both grant-based and contention-based access protocol enhancements. Grant-based allocations examine diversity and interference limiting techniques to improve availability, and contention-based access schemes ensure reli-

ability by bringing down collisions. In the next section, we present allocation designs inspired by grant-based [PIV, PV, PVI] and contention-based [PVII] allocations that improve 5G URLLC availability and reliability.

4.1 Grant-based access

In this section, we concentrate on URLLC availability in a multi-UE, multi-cell network. We have proposed solutions in [PIV, PV PVI] using multi-hop relaying and multi-point coordination schemes to reduce loss margins and restrict interference, thus improving the communication availability.

4.1.1 Multi-hop access

We consider a cooperative relay network with multiple RNs to choose from. RNs are assumed to perform CSI estimation for the hop and report results to gNB. The gNB estimates the UE rate. Selective multi-hop relaying is applied to maximize the rate available to 99.999 % of UEs. The gNB applies a threshold, and if the direct link rate of a UE is below this threshold, multi-hop communication is considered. The gNB selects RN that provides the highest achievable transmission rate. Two modes, (a) Half-duplex and (b) Full-duplex communication, are considered.

Half-duplex relaying

In half-duplex relaying, gNB i transmits data to RN j in the first hop-period; RN then decodes and retransmits data to UE k in the next hop. These hops are orthogonal in time and thus generate no intra-flow interference. The e2e rate in half-duplex transmission is $r_{ik} = \min\{\tau_{ij}\rho_{ij}, \tau_{jk}\rho_{jk}\}$ where τ_{ij} and τ_{jk} are time scheduling weights, and ρ_{ij} and ρ_{jk} are rates observed for a unit time hop-period. If sub-optimal equal hop-periods $\tau_{ij} = \tau_{jk} = \tau_{ik}/2$ are used, the transmission rate is $r_{ik} = \frac{\tau_{ik}}{2}\min\{\rho_{ij}, \rho_{jk}\}$. The rate maximization problem for identifying hop-period weights is

$$\begin{aligned} \underset{\tau_{ij},\tau_{jk}}{\text{Maximize}}: \quad & r_{ik}, \\ \text{Subject to}: \quad & \tau_{ij} + \tau_{jk} = \tau_{ik}, \\ & \tau_{ij} > 0,\ \tau_{jk} > 0. \end{aligned}$$

The solution to the problem delivers equal rates for both hops, i.e., $\tau_{ij}\rho_{ij} = \tau_{jk}\rho_{jk}$, where $\tau_{ij} = \tau_{ik}\frac{\rho_{jk}}{\rho_{ij}+\rho_{jk}}$ and $\tau_{jk} = \tau_{ik}\frac{\rho_{ij}}{\rho_{ij}+\rho_{jk}}$ [PV].

Full-duplex relaying

In full-duplex relaying, RN receives and transmits simultaneously. The full-duplex operation produces both loop-back self-interference and forward interference. Self-

interference occurs when the transmission of a node interferes with its reception. With state-of-the-art full-duplex designs [PV], self-interference may be curtailed to a large extent. Forward interference arises when the transmission on the previous hop of flow interferes with the other hop's reception, e.g., node i's transmission interferes with the reception of transmission from node j at destination node k.

Power control optimization

Intra-flow power control optimization, similar to [96], can be applied, where each gNB employs power control independently at gNB/RN to suppress forward and self-interference. Simplified power control optimization problem can be formulated based for e2e transmission rate maximization as

$$\begin{aligned} \underset{p_i, p_j}{\text{Maximize}}: \quad & r_{ik}, \\ \text{Subject to}: \quad & 0 < p_i \leq p^{\max}, \\ & 0 < p_j \leq p^{\max}. \end{aligned}$$

The solution to the above optimization problem is to keep maximum power on the bottleneck hop and reduce power on the superior hop so that $r_{ij} = r_{jk}$. The superior hop is the one that has a higher rate when $p^{\max}$ is used for both hops. Fixing transmit power of the weaker hop to be $p^{\max}$, the power of the superior adjusted to p_s that maximizes the net rate. The power optimization problem always has a unique solution in the range $p_s \in [0, p^{\max}]$, which can be found, e.g., by line search. An example solution is presented in [PV].

Coordinated precoding optimization

For MIMO communication systems, coordinated precoding based optimizations can be aimed for rate maximization. For example, in [97], spatial directions of two transmissions, i.e., precoders, are jointly optimized to mitigate interference, which assumed a joint power budget for gNB and RN to formulate an optimization problem for subcarrier allocation. Here, the total transmit power is not constrained in a network, and we assume that hardware restrictions only limit the maximum transmission power of different nodes, and precoders are jointly optimized to circumvent residual self-interference.

The gNB optimizes precoding matrices $\mathbf{W}_{ij}$, $\mathbf{W}_{jk}$ using the conjugate gradient search method to cause its hop rates r_{ij}, r_{jk} to become close together. Precoding

matrices are updated accordingly

$$\mathbf{W}_{ij} = \begin{cases} \mathbf{W}_{ij} + \epsilon \frac{dr_{jk}}{d[\overline{\mathbf{W}}_{ij}]}, & \text{if } r_{ij} > r_{jk} \\ \mathbf{W}_{ij} + \epsilon \frac{dr_{ij}}{d[\overline{\mathbf{W}}_{ij}]}, & \text{if } r_{ij} < r_{jk} \end{cases},$$

$$\mathbf{W}_{jk} = \begin{cases} \mathbf{W}_{jk} + \epsilon \frac{dr_{jk}}{d[\overline{\mathbf{W}}_{jk}]}, & \text{if } r_{ij} > r_{jk} \\ \mathbf{W}_{jk} + \epsilon \frac{dr_{ij}}{d[\overline{\mathbf{W}}_{jk}]}, & \text{if } r_{ij} < r_{jk} \end{cases}$$

where ϵ is a small step size.

We consider numerical analysis focusing on maximizing the rate available to 99.999 % of UEs by applying selective multi-hop relaying. Multi-hop relaying is selectively used to the worst 0.1 % of UEs. With multi-hop optimization schemes, UEs in outage can enhance their QoS multi-folds and become part of the high-capacity region, i.e., upper tail of CDF. The precoding weights optimization brings the most gains. Gains are depicted in Fig. 4.1. Simulation parameters and thorough analysis are detailed in [PV]. Gains at 100 % of CDF are recorded over the range of 5000-8000 %, with full-duplex communication performing best. However, full-duplex relaying implementation is complex due to its physical design and optimization issues, exhibiting a complexity-gain trade-off. When multi-hop is extended to all UEs, the rates in the ultra-reliable communication region diminish due to increased interference from additionally activated RNs in contrast to single-hop and are depicted in [PV].

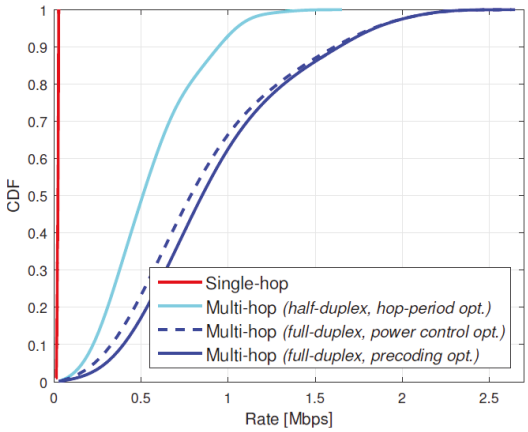

Figure 4.1. Rate CDF improvement of single-hop low rate UEs via varying multi-hop schemes [PV].

4.1.2 Interference-aware scheduling

In a multi-cellular environment, URLLC can be critical to inter-cell interference. To observe interference impact, various layouts of an isolated factory with different

gNB density are simulated in [PIV]. The related UE rate distributions are depicted in Fig. 4.2. It appears that the UE rate available at 99.999 % target (10^{-3} % of CDF) for the case with single gNB outperforms others, shown with a zoom-in view in Fig. 4.2. It shows that inter-cell interference can be highly damaging to URLLC in a multi-cellular network. It should be noted that here we consider an isolated factory, and there is no external interference. In the presence of external interference, e.g., from outside macro gNB, the QoS can degrade, and the relative gains for a single gNB scenario may not be as high. Further, we utilize two primary solutions — power control and orthogonal sharing to limit interference in a multi-cell network [PIV]. Undoubtedly, orthogonal sharing is extremely good in eliminating interference, but it comes with a price of lower spectrum usage. Orthogonal sharing brings the highest gains in the ultra-reliable region, but UEs capacity performance suffer due to reduction in BW resources. On power control optimization in a full sharing setting between gNBs, gains are negative compared to a single gNB network. It shows that we need better coordination techniques among neighboring cells to restrict interference and simultaneously operate with maximum frequency re-use.

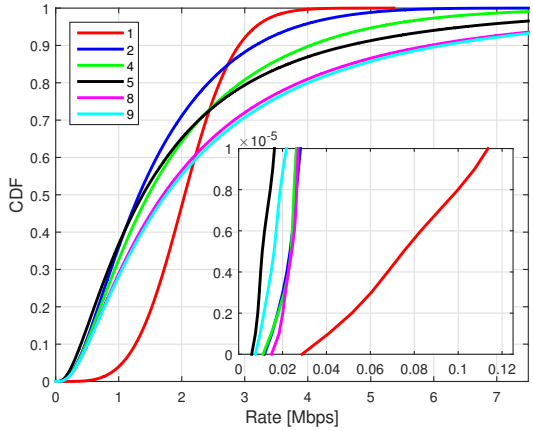

Figure 4.2. UE rate distribution in factory layout equipped with different numbers of gNBs. Zoom-in to the URLLC region around 99.999 % availability [PIV].

Having multiple gNBs deployment in a geographical area has its advantages. It can dramatically improve coverage or availability. In a single cell network, gNB failure can cause a 100 % network outage when all machines rely on a single gNB. Besides, packet delays and jitter are significantly decreased when there are multiple gNBs due to smaller wireless propagating distances. On the other hand, it causes inter-cell interference and thus, we explore a scheduling design that improves URLLC service by limiting interference. We consider a max-min scheduling design that enhances the minimum rate UE in a complete sharing setting for multi-cell MIMO network [VI]. The scheduling operation requires coordination between gNBs to construct and ex-

change Background Interference Matrices (BIMs). Using BIMs information, gNBs ensure that while implementing max-min scheduling operation, they do not increase interference to victim UEs of the opponent gNBs, thereby improving the minimum network throughput. This type of coordination is reasonable where competitive nodes are servicing URLLC UEs at the cell edge. Instead of serving UEs with full power and inflicting heavy interference, these nodes can coordinate in a limited manner while scheduling URLLC UEs at cell edge where interference can be significant. They are only required to exchange information related to BIMs and partial information about scheduling pattern of cell-edge URLLC UEs. This type of behavior contrasts with spectrum allocation protocols discussed in Chapter 3, where no information is exchanged among competitive nodes. The scheduling protocol is divided into two steps and is described as following.

Construction and exchange of the BIMs

The gNBs coordinate, construct and exchange BIMs. A BIM contains all possible interference values a UE can perceive from the opponent gNB's set of precoded signals. In order to construct BIM, first, one of the nodes, let us say gNB A will send the request to the opponent gNB B for BIM construction and its reporting. In response, gNB B will send an acknowledgment and a broadcast message to its UEs to report inference powers perceived by them due to undesired signals from gNB A. Once gNB B fetches all these estimated interference power levels from its UEs, it then forwards interference measurements to gNB A. Hence, BIM for gNB A is represented as

$$\mathrm{BIM}_A = \begin{bmatrix} I_{A_1,1} & \cdots & I_{A_1,y} \\ \vdots & \ddots & \vdots \\ I_{A_x,1} & \cdots & I_{A_x,y} \end{bmatrix} \tag{4.1}$$

where $I_{A_x,y}$ is interference power from gNB A for precoded signal meant for its UE x to the opponent gNB B's UE y.

Iterative max-min scheduling

After exchanging BIMs, each gNB knows the interference power it can inflict while allocating its minimum rate UE additional resources (symbols). Hence, only those symbols are considered in which the gNB does not increase interference from the current value to the opponent UE. Each gNB iteratively runs the scheduling algorithm and continues to do so until the minimum throughput stops increasing. The optimization problem for max-min optimization is detailed in [PVI]. Essentially, in this max-min scheduling process, the gNB checks the interference power it inflicts on the victim UE. The interference power is deduced from the information gNB possesses — a scheduling pattern of the opponent cell indicating victim UEs and the mapped interference power to victim UEs from BIM. While allocating its minimum

rate UE an additional symbol, if the aggregate interference power does not increase to the victim UE, only then scheduling decision is allowed. In this manner, gNBs try to maximize their minimum rates, and at the same, they do their best not to increase interference power to victim UEs.

To assess the performance of coordination strategy, the lower tail of UE rate CDF is compared with max-min scheduling operation but without coordination for availability improvement at 99.999 % target. Simulation parameters are detailed in [PVI]. The interference-aware scheduling is selectively applied to UEs with achievable default rates located in the worst 0.1 % distribution tail. In Fig. 4.3, we depict outage UEs with and without a coordination scheme. The outage is reduced by below 50 % at 99.999 % availability target by employing the coordination scheme. UEs in the outage with poor achievable rates are scheduled on resources with less interference. This results in improvement of rates and diminishes outage probability in the network.

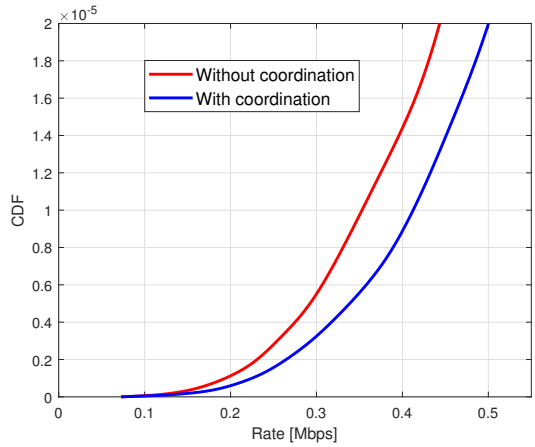

Figure 4.3. Rate distribution for outage UEs with and without coordination. The coordination protocol enables gNBs to maximize their minimum rate without increasing inter-cell interference on the common resource compared to no coordination scenario [PVI].

4.2 Grant-free access

For low latency uplink access, grant-based access schemes can incur moderate to significant latency due to control signaling in the form of SR, Buffer Status Report (BSR), and subsequent grant allocations. Moreover, there is an associated probability of decoding error of uplink and downlink control signaling, thus increasing the PER.

An uplink GF, contention-based access scheme could be a viable alternative for URLLC services. The UE transmits its uplink data in an arrive-and-go manner without sending SR and receiving resource grants from the network. It reduces latency, and therefore, it has a growing interest among notable industrial players for 5G

URLLC services [98]. GF transmission can be done in a dedicated [99] or shared resource [PVII,100]. However, allocating dedicated resources for sporadic URLLC traffic would lead to the wastage of costly spectrum resources. Thus, we are interested in GF, contention-based access for URLLC on a shared resource, and collisions may occur among different UEs. To further reduce latency, we consider NR numerology of reduced slot size of 0.125 ms. With this, on average, 4 slots are available for transmission in one direction with 0.5 ms latency.

The reliability of contention-based URLLC transmission can degrade quickly due to collisions among competing UEs for channel access. Hence, we consider diversity-based solutions [PVII], similar to [101] to improve the reliability of multi-channel slotted ALOHA. In [PVII,101], the same data packet from a UE is sent multiple times in consecutive slots. As a result, the packet success rate is improved at the expense of transmission redundancy still within the tight URLLC latency window, even if some repetitions end up in collisions. In contrast to [101], we deduce the optimal URLLC load or diversity count utilizing approximate modeling for sporadic arrivals.

We consider allocation design where GF transmission's collision probability can be reduced by employing diversity transmission or repetition coding. In this scenario, a UE transmits the same data packet in Γ subsequent slots where each repetition requires a unit RB in frequency domain. However, in every subsequent slot, the RB for packet transmission is chosen randomly from K available RBs. As a result, some packets may collide, but some may be transmitted successfully. With Γ diversity transmission, each UE transmits its data packet Γ times in Γ consecutive slots. We treat K contention resources in Γ slots as a new contention pool, where UE chooses resource from K^{Γ} resources and the traffic intensity ψ is increased to $\Gamma\psi$. From collision probability definition in (2.5), the new collision probability with diversity transmission then becomes

$$P_{\Gamma}^{\text{col}} = 1 - \left(\frac{e^{-\Gamma\psi} + K^{\Gamma} - 1}{K^{\Gamma}} \right)^{N-1}. \tag{4.2}$$

Table 4.1 details diversity levels and corresponding supported UE load sizes for different arrival rates, estimated based on Eq. (4.2) at 99.999 % reliability target which is $1 - P_{\Gamma}^{\text{col}}$. The transmission parameters are detailed in [PVII]. It can be concluded that at sporadic arrivals, higher diversity substantially increases support for URLLC load, and vice-versa, and the supported load starts diminishing with an increase in arrival rate.

Next, we present performance curve depicting collision rate CDF against the supported load. In addition to transmission diversity, we further analyze multi-UE detection and combined technique utilizing both multi-UE detection and diversity transmission to measure performance. Multi-UE detection allows detection of collided

Table 4.1.
URLLC UE load sizes for arrival rates and diversity degrees. Contention pool with $K = 6$ RBs.

UE population (N)		Diversity (Γ)			
		1	2	3	4
Mean	1.25×10^{-6}	49	145	577	2593
arrival	1.25×10^{-5}	5	15	58	260
rate ψ	1.25×10^{-4}	1	2	6	26

packets with some degree of success and is modeled in [PVII]. In Fig. 4.4, the supported load is plotted against various techniques at a mean arrival rate of 1.25×10^{-5}. The maximum supported load with alone diversity and combined technique is 260 UEs and 428 UEs against five UEs with default ALOHA-based transmission at 99.999 URLLC reliability target. The achievable gains with combined technique is 8500 %. This shows that collisions lower dramatically with diversity transmissions, and gains are further supplemented with multi-UE detection techniques. In addition, a case with flexible access is depicted, the allocation design allows UEs to transmit on dedicated $K = N$ CCs if value K derived from Eq. (4.2) is more than N, i.e., $K > N$. For example, in Table 4.1, we see a load less than six UEs still require $K = 6$ CCs in order to limit collision probability below 0.00001. Each UE can be scheduled a dedicated CC in such scenarios and thus resource usage would incur less than 6 CCs with 0 collision probability.

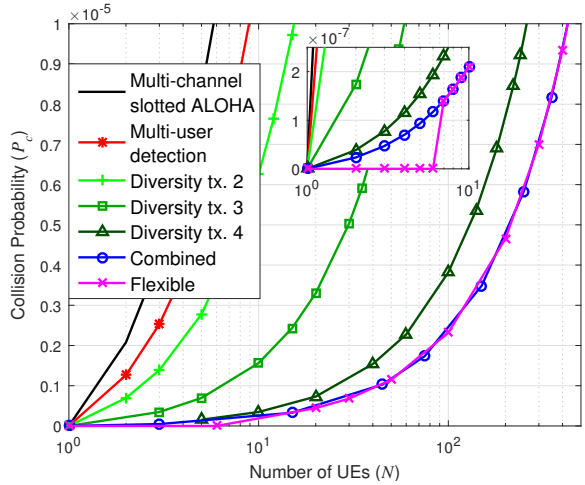

Figure 4.4. Collision rate CDF against the URLLC load. The zoom-in part depicts the collision rate at the bottom of the CDF tail [PVII].

The discussed model here presumes sporadic transmissions, and the model is approximated by ignoring two events — (a) self-collisions and (b) collisions among UEs with different packet replicas. By self-collisions, we mean the UE's multiple packets arrive in the same access slot and may collide with each other. However, ignoring self-collision events makes sense at low arrival rates as the probability of multiple packets arriving in the same access slot is negligible. Moreover, even if they arrive simultaneously, in practice, the UE has absolute control over transmissions. The UE is fully capable of prohibiting collisions by allowing packets' transmissions as per queue standing in the buffer. On the other hand, the effect of collisions between different replicas can be more prominent. An example with $\Gamma = 2$ would be that a UE's first repetition collides with one UE and its second repetition collides with a different UE. When ignoring such events we can compute the number of resource K or the number of supported users N in closed-form for a given reliability target based on Eq. (4.2). However, the approximate collision probability (4.2) breaks down at larger intensities ψ. This is visible in the results of [PVII], where we see that the analytical model breaks down at a high arrival rate. Moreover, the higher the diversity degree, the earlier the model breaks. To assess the approximation error at low arrival rates, we compare the approximation in [PVII] with analytical upper and lower bounds of the true expression. The collision upper bound is estimated by assuming a synchronized UE arrival pattern; if UEs are configured with Γ repetitions, we assume that UE packets can arrive only at access slots $\varsigma\Gamma$ where ς is an integer. It means that if $\Gamma = 4$, UE packets can arrive at access slots $n = 0, 4, 8, \ldots$ The collision upper bound is

$$P_{\Gamma}^{\text{col,upper}} = \sum_{n=1}^{N-1} \frac{N-1!}{N-n-1!\,n!} \Big(1 - e^{-\Gamma\psi}\Big)^{n} \Big(e^{-\Gamma\psi}\Big)^{N-n-1} \Bigg(1 - \bigg(\frac{K-1}{K}\bigg)^{n}\Bigg)^{\Gamma}. \quad (4.3)$$

To calculate the lower bound, we assume that the Γ repetitions are not consecutive, rather fully independent. Thus, the lower collision bound is derived from Eq. 4 in [PVII] as $\left(P_c\left(\Gamma\psi\right)\right)^{\Gamma}$, i.e.,

$$P_{\Gamma}^{\text{col,lower}} = \left(1 - \left(\frac{e^{-\Gamma\psi} + K - 1}{K}\right)^{N-1}\right)^{\Gamma}. \quad (4.4)$$

Results for approximated model in [PVII] are plotted in Fig. 4.5, together with the upper and lower bounds (4.3) and (4.4). It can be seen that at low arrival rates, which is our focus of interest, collision rate interestingly interpolates between upper and lower bounds and lies closer to the upper bound. This shows that by discounting events mentioned above and assuming simplistic modeling as in Eq. (4.2), the collision rate is close to the upper bound. This means that actual collision probability based on accurate modeling can be lower than (4.2), i.e., even more UEs can be supported for the given resource and reliability target.

Scenarios with a high population or high arrival rate may not benefit from diversity

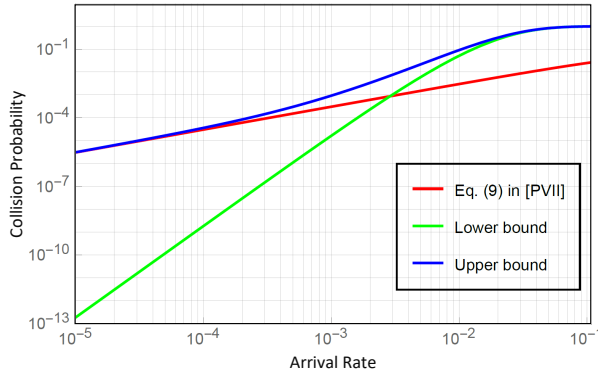

Figure 4.5. Collision rate CDF plotted for [PVII] model; the upper bound based on Eq. (4.3) and the lower bound based on Eq. (4.4).

transmission. They are effective only in a relative abundance of resources. Thus, for UEs with a high arrival rate, scheduled or dedicated GF access is desirable. The reason is that the whole contention resource K may run into outage due to many packets and their replicas, discussed in [PVII]. In conclusion, URLLC services with significant traffic cannot be established without scheduled access.

4.3 Summary

In this chapter, the resource allocation for critical services is discussed. These services require high availability, high reliability, and low latency. To provide extreme availability, solutions based on transmission link diversity are discussed by deploying half-duplex and full-duplex relay nodes to enable multi-hop communication. Multi-hop relaying is applied selectively to the worst rate users in a geographical area, and the users boost their rates considerably by choosing new routes with less shadowing and inter-cell interference. We have considered optimization techniques to maximize the net rate over multi-hop by making rates on each hop approximately equal. For instance, in half-duplex relaying, we optimize hop-period such that the hop with poor rate is allocated with relatively longer time resource, and vice-versa. In full-duplex relaying, we optimize power or precoder for each hop which limits intra-flow and self-interference and maximizes the net transmission rate.

In addition, we have discussed an optimization scheme to control inter-cell interference. The neighboring cells coordinate and construct background interference matrices where each cell records interference powers the other cell can inflict for the countable set of precoders. The cells further coordinate to perform iterative scheduling where they exchange limited scheduling information and coordinate to perform max-min scheduling. They make sure that the updated scheduling patterns do not in-

crease interference powers from previous values to victim users by examining them from background matrices. Both cells continue to perform scheduling iterations until the minimum throughput cannot be increased.

To provide extremely low latency, contention-based schemes could be aimed at users with sporadic traffic. The user transmits a packet in a grant-free manner without requiring to send a scheduling request or waiting for a grant message. If two or more users attempt to transmit simultaneously over the shared resource, this can result in collisions. To improve the reliability against collisions, diversity transmission can be utilized where packet replicas are sent multiple times to reduce the collision probability. These complex interactions involving diversity transmission are constructed with an approximate model tailored for sporadic traffic, which helps to identify shared resource pool size or repetition count with simplicity. When the traffic intensity is increased, the model starts deviating from simulated behavior and thus limiting its applicability to sporadic traffic.

5. Conclusion and Future Work

5.1 Conclusion

In this thesis, resource allocation for 5^{th} generation systems is discussed, aiming to deliver efficient mobile broadband and ultra-reliable, low latency communication services. To enable efficient broadband services, a spectrum sharing framework with a non-cooperative game model is presented. The two-player game is formulated, which requires no revelation of player's critical information to each other or any third-party entity. The formulated game is repeated and provides gain relative to a one-shot game or no sharing scenarios. The outcome of a cooperative game act as an upper bound for the non-cooperative game. This is as expected; the gains from cooperation come at the price that operators must reveal their critical information like traffic or load information. Hence, due to this demanding requirement, competitive operators are not likely to enter into a cooperative game, due to lack of trust. The discussed non-cooperative game allows operators to determine their actions independently. They are free to act in their self-interest, e.g., they can deny opponent operator's spectrum requests or always demand spectrum favors. However, the gains will become only prominent if both operators behave rationally, i.e., exchanging favors when one has a more dire need, and the game converges to Nash equilibrium.

The latter half of the thesis considers the investigated critical communication framework employing grant-based and grant-free access. To provide extreme availability among scheduled users, solutions employing diversity become essential, e.g., by enabling multi-hop schemes to provide availability gains. The multi-hop relaying is optimized where the net rate over two hops is maximized by adjusting hop-period, powers, or precoders. Further, if neighboring cells are interfering, solutions must be considered to counter or limit interference. In one of the demonstrated solutions, neighboring cells coordinate to design precoders and thus improve their availability. In order to provide extremely low latency, contention-based grant-free access is con-

sidered where users transmit readily on the shared resource pool, and collisions may occur between transmissions from different users. The problem explores resource allocation optimization techniques for users with sporadic traffic. Solutions such as multi-user detection and multiple repetitions are considered to limit collisions, thus improving reliability. In New Radio, an allocation based on configured grant is standardized where users also transmit readily on their pre-configured resources [31]. These resources can overlap between users, and as a result, collisions can happen. The previously discussed solutions for contention-based access can be applied here to limit collisions over overlapping configured grants. Hence, solutions based on a shared resource for low latency critical communication could be a cost-effective means, which this thesis has tried to explore.

5.2 Future work

The work in this thesis can be extended in numerous ways. The non-cooperative game model can be expanded to multiple operators sharing spectrum. This will increase the possibility of securing a favor due to the increased number of players, but at the same time, it makes the problem of finding an optimal or Nash equilibrium state complex. In this thesis, operators with similar behavior are considered, e.g., operators implementing the same utility function and the same contribution towards spectrum pool. Thus we address the research question raised in Chapter 1 for specific operator behavior. In future work, the analysis can be extended to games among operators with different behaviors to improve sharing capacity, e.g., operators may employ different utility functions, contribute a different amount of spectrum towards the pool, or have an unequal mean load. Such techniques can become prominent, e.g., in spectrum band 52.6 to 71 GHz, which is currently being standardized in New Radio Release 17 [102], or in prospective THz bands for future generations. The spectrum resource in higher frequencies does not interfere strongly due to innate characteristics such as higher path loss and narrow beamforming. Hence, for future work, the analysis can be expanded to high-frequency channel models, and the likely enhancements can be explored, given that residual interference inflicted by the neighboring network may not be severe.

To improve critical services, we have discussed solutions such as multi-hop communication, coordination to limit external interference, and repetition coding. Thus, the question put up in Chapter 1 concerning ultra-reliable services is answered in part. Two main conclusions can be deduced. First, diversity in some form is important to improve reliability or availability. Second, inter-cell coordination becomes necessary when networks are indulging in frequency reuse techniques. With ongoing efforts

in New Radio standardization, repetition coding has been implemented to improve reliability. For low latency access, 5th generation New Radio has standardized configured grant allocation, a type of periodic resource allocation, which can be activated by transmitting one-time radio resource control or downlink control signaling [31]. The performance benefits can only be observed in a contention-free environment, as there are no collision resolution techniques standardized. Hence, for future work, the research question can be further investigated for hybrid scenarios where some repetitions can be done exclusively and the remaining repetitions in a contention-based manner. Depending on load, reliability, latency requirements, and channel characteristics, the allocation can be optimized to provide dedicated, contention-based, or hybrid allocation. The hybrid solution can be considered as utilization of both New Radio and New Radio Unlicensed.

Another area of interest is the utilization of channel sensing mechanisms to avoid collisions. New Radio Unlicensed has implemented the listen-before talk channel access mechanism for collision avoidance in the license-exempt spectrum. For example, in frame-based equipment or semi-static access mode [19], the user senses the channel for 9 µs before the scheduled transmission given that the transmission begins after the idle period or the transmission gap is greater than 16 µs. If the sensed channel is not free, the user drops the transmission. This helps users or nodes associated with different cells or Wi-Fi access points to avoid transmitting simultaneously on the common resource. Extending the listen-before talk mechanism to avoid collisions among users competing for resources within the New Radio or New Radio Unlicensed cell can offer benefits over collision resolution techniques which utilize more resources in the form of repetitions or retransmissions. However, the problem is that due to the slotted nature of the resource, the users within the cell will begin at the same slot or symbol boundary assuming perfect timing advance implementation. This means the existing sensing will not help to avoid collisions. This would require a new type of channel access design for users competing for resources within the same cell to avoid collisions. Further, the license-exempt resource use gives arise to two-tier architecture, where both intra-cell and inter-cell collisions can occur. This offers an interesting problem scenario that necessitates a solution design by possibly integrating multiple channel access mechanisms.

Interference mitigation is extremely important for ultra-reliable service provisioning, and this is discussed in [PIV, PVI]. 5th generation New Radio has standardized solutions to handle intra-user and intra-cell interference in case of multiplexed services. Solutions, such as deprioritization, cancellation, or preemption prioritize ultra-reliable transmissions, and the overlapping mobile broadband transmissions are canceled. The degradation of capacity performance can be noticeable due to cancel-

lations as mobile broadband transmissions have relatively larger transport block size, and their retransmissions can cause resources wastage. These standardized solutions are elementary, which only focus on protecting critical transmissions, and are not optimal for providing overall capacity performance. This could be an another avenue for future research or communication design for improved interference handling with multiplexed services.

Further, a possible harmonization of two topics of the thesis can be considered for immersive services such as extended reality [13]. Such services require high capacity and high reliability within a bounded latency. In other words, it can be regarded as a type of broadband ultra-reliable and latency critical service, and is therefore extremely BW hungry. This thesis has provided optimization techniques to separately consider broadband and ultra-reliable and latency critical scenarios. In future work, a joint problem related to resource allocation for extended reality services can be considered where solutions can be built or improved by harmonizing the solutions presented in this thesis.

With ongoing standardization of the 5^{th} generation New Radio and continuing network roll-out, 6^{th} generation research has already started both in industry and academia [103]. 5^{th} generation standardization has commenced supporting new verticals and use cases, and it is believed that its successor will carry onto this trend and support an explosive number of uses cases, such as support for machine learning-driven networks or digital twins [103, 104]. 3GPP has standardized features for machine-type communications which enable ultra-reliable and low latency services. The adoption of these features for local industrial networks will take few years, and the confronted issues will provide inputs to 6^{th} generation research and standardization efforts. One particular area of interest is the artificial intelligence and machine learning components in 5^{th} generation system, which are still in their infancy and are part of the scope in study items [105, 106]. These technology components are expected to be fundamental blocks in 6^{th} generation networks enabling intelligent traffic control [103]. For instance, the artificial intelligence and machine learning integration can automate user scheduling. It can enable seamlessly switch between the dedicated, contention-based, or hybrid access adapting at the medium access control layer to the traffic and channel requirements. The next generation is also envisioned to utilize THz bands. This may introduce changes to the physical layer design due to the extremely small slot size and communication channels characterized by large path loss.

References

[1] Ericsson, “Mobile data traffic outlook," Jun. 2019. [Online]. Available: https://www.ericsson.com. [Accessed: Aug. 19, 2019].

[2] ITU-R, “IMT vision - Framework and overall objectives of the future development of IMT for 2020 and beyond," Rep. M.2083-0, Sep. 2015.

[3] ABI, “Planned cost efficiencies and support for a wider range of automotive use cases to drive adoption of C-V2X technology," Feb. 2018. [Online]. Available: https://www.abiresearch.com. [Accessed: Aug. 19, 2019].

[4] ITU-R, “IMT traffic estimates for the years 2020 to 2030," Rep. M.2370-0, May, 2015.

[5] CCS Insight, “Market forecast 5G connections, worldwide 2018-2025," Rep., Aug. 2018.

[6] Cisco, “Global mobile networks will support more than 12 billion mobile devices and IoT connections by 2022; Mobile traffic approaching the Zettabyte milestone," Feb. 2019. [Online]. Available: https://newsroom.cisco.com. [Accessed: Aug. 29, 2020].

[7] S. E. Elayoubi *et al.*, “5G service requirements and operational use cases: Analysis and METIS II vision," in *Proc. EuCNC*, pp. 158-162, Jun. 2016.

[8] Qualcomm, “The evolution of mobile technologies: 1G; 2G; 3G; 4G LTE," Jun. 2014. [Online]. Available: https://www.qualcomm.com. [Accessed: Aug. 19, 2019].

[9] 3GPP, “Study on scenarios and requirements for next generation access technologies," TR 38.913 V15.0.0, Jun. 2018.

[10] Qualcomm, “Making 5G NR a reality," White paper, Dec. 2016.

[11] 3GPP, “Study on communication for automation in vertical domains," TR 22.804 V16.2.0, Dec. 2018.

[12] N. Xia, H. Chen and C. Yang, “Radio resource management in machine-to-machine communications - A survey," in *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 791-828, First quarter 2018.

[13] 3GPP, “Technical specification group services and system aspects; Extended Reality (XR) in 5G (Release 16)," TR 26.928 V16.1.0, Dec. 2020.

[14] Č. Stefanović, “Industry 4.0 from 5G perspective: Use-cases, requirements, challenges and approaches," in *Proc. CMI Int. Conf.*, pp. 44-48, Nov. 2018.

[15] 3GPP, “38 series.” Available: `https://www.3gpp.org/DynaReport/38-series.htm`. [Accessed: Sep. 07, 2020].

[16] 3GPP, “Physical channels and modulation," TS 38.211 V15.6.0, Jun. 2019.

[17] 3GPP, “Base station (BS) radio transmission and reception," TS 38.104 V16.0.0, Jun. 2019.

[18] 3GPP, “Study on NR-based access to unlicensed spectrum," TR 38.889 V16.0.0, Dec. 2018.

[19] 3GPP, “Physical layer procedures for shared spectrum channel access," TS 37.213 V16.6.0, Jun. 2021.

[20] 3GPP, “System architecture for the 5G System (5GS); Stage 2," TS 23.501 V17.1.1, Jun. 2021.

[21] 3GPP, “Study on physical layer enhancements for NR ultra-reliable and low latency case (URLLC)," TR 38.824 V16.0.0, Mar. 2019.

[22] 3GPP, “Study on NR Industrial Internet of Things (IoT)," TR 38.825 V16.0.0, Mar. 2019.

[23] 3GPP, “Physical channels and modulation" TS 36.211 V15.6.0, Jun. 2019.

[24] 3GPP, “Physical layer procedures for control," TS 38.213 V15.6.0, Jun. 2019.

[25] 3GPP, “NR; User equipment (UE) radio transmission and reception; Part 2: Range 2 standalone," TS 38.101-2 V16.0.0, Jun. 2019.

[26] 3GPP, “NR; User equipment (UE) radio transmission and reception; Part 1: Range 1 standalone," TS 38.101-1 V16.0.0, Jun. 2019.

[27] J. Bae *et al.*, “An overview of channel coding for 5G NR cellular communications," *APSIPA Trans. Signal Inf. Process.*, vol. 8, pp. 1-14, Jun. 2019.

[28] A. E. S. Hassan *et al.*, “Evaluation of complexity versus performance for Turbo code and LDPC under different code rates," in *Proc. SPACOMM*, pp. 1-6, Apr. 2012.

[29] 3GPP, “Physical layer procedures" TS 36.213, V16.2.0, Jun. 2020.

[30] 3GPP, “Packet Data Convergence Protocol (PDCP) specification" TS 38.323, V16.1.0, Jul. 2020.

[31] 3GPP, “Physical layer procedures" TS 38.214, V16.2.0, Jun. 2020.

[32] M. Gerami and B. Singh, “Configured Grant for Ultra-Reliable and Low-Latency Communications: Standardization and Beyond," submitted in IEEE Commun. Standards Mag., 2020.

[33] Z. Li *et al.*, “5G URLLC: Design challenges and system concepts," in *Proc. ISWCS*, pp. 1-6, Aug. 2018.

[34] O. Iscan, D. Lentner and W. Xu, “A comparison of channel coding schemes for 5G short message transmission," in *Proc. IEEE Globecom Workshop*, pp. 1-6, Dec. 2016.

[35] S. Tabbane, “IoT long range technologies: Standards," ITU, Dec. 2017.

[36] N. H. Mahmood *et al.*, “Reliability oriented dual connectivity for URLLC services in 5G New Radio," in *Proc. ISWCS*, pp. 1-6, Aug. 2018.

[37] 3GPP, “Final Report of 3GPP," TSG RAN WG1 #97 V1.0.0, May 2019.

[38] N. Attey, S. Maximov and M. A. El-Saidny, “5G NR; A new era for enhanced mobile broadband," Mediatek, White paper, 2018.

[39] C. E. Shannon, “Communication in the presence of noise," in *Proc. IRE*, vol. 37, no. 1, pp. 10-21, Jan. 1949.

[40] C. Walck, *Hand-book on statistical distributions for experimentalists*, Particle Physics Group, Fysikum, University of Stockholm, Sep. 2007.

[41] E. Dahlman, S. Parkvall and J. Sköld, *5G NR: The next generation wireless Access technology*, Academic Press, 2018.

[42] N. Patriciello *et al.*, “NR-U and IEEE 802.11 technologies coexistence in unlicensed mmWave spectrum: Models andeEvaluation," *IEEE Access*, vol. 8, pp. 71254-71271, 2020.

[43] F. Kelly, “Charging and rate control for elastic traffic," *European Trans. Telecomm.*, vol. 8, pp. 33-37, 1997.

[44] C. W. Sung and W. S. Wong, “A noncooperative power control game for multirate CDMA data networks," *IEEE Trans. Wireless Commun.*, vol. 2, no. 1, pp. 186-194, Jan. 2003.

[45] B. Singh, Z. Li and M. A. Uusitalo, “Flexible resource allocation for device-to-device communication in FDD system for ultra-reliable and low latency communications," in *Proc. RTUWO*, pp. 186-191, Nov. 2017.

[46] K. Navaie, D. Montuno and Y. Q. Zhao, “Fairness of resource allocation in cellular networks: A Survey," in *Resource allocation in next generation wireless networks*, Nova Science Publishers, New York, 2005.

[47] X. Gao, T. Nandagopal and V. Bharghavan, “Achieving application level fairness through utility-based wireless fair scheduling," in *Proc. IEEE GLOBECOM*, pp. 3257-3261, Nov. 2001.

[48] W. Xu and J. Wang, “Double auction based spectrum sharing for wireless operators," in *Proc. IEEE PIMRC*, pp. 2650-2654, Sep. 2010.

[49] F. P. Kelly, A. K. Maulloo and D. K. H. Tana, “Rate control for communication networks: Shadow prices, proportional fairness and stability," *J. Operational Research Soc.*, vol. 49, no. 3, pp. 237-252, 1998.

[50] J. Mo and J. Walrand, “Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556-567, 2000.

[51] C. U. Saraydar, N. B. Mandayam and D. Goodman, “Efficient power control via pricing in wireless data networks," *IEEE Trans. Commun.*, vol. 50, no. 2, pp. 291-303, Feb. 2002.

[52] S. Keshav, *An engineering approach to computer networking*, Addison-Wesley, Reading, MA, 1997.

[53] J. Huang, R. A. Berry and M. L. Honig, “Auction-based spectrum sharing," *Springer Mobile Netw. Applicat.*, vol. 11, no. 3, pp. 405-418, Jun. 2006.

[54] B. Singh *et al.*, “Resource allocation strategy for ultra-reliable communication in a factory environment," in *Proc. Golbal Wireless Summit*, pp. 1-6, Nov. 2016.

[55] N. Abramson, “The ALOHA system - Another alternative for computer communications," in *Proc. AFIPS Fall Joint Comput. Conf.*, vol. 37, 1970.

[56] L. G. Roberts, “ALOHA packet system with and without slots and capture," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 5, no. 2, pp. 28-42, Apr. 1975.

[57] 3GPP, “Study on new radio access technology physical layer aspects," TR 38.802 V14.0.0, Tech. Rep., Sep. 2017.

[58] A. Apostolidis *et al.*, “Intermediate description of the spectrum needs and usage principles,” METIS, Rep. ICT-317669-METIS/D5.1, 2013.

[59] A. Osseiran *et al.*, “Scenarios for 5G mobile and wireless communications: The vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26-35, May 2014.

[60] T. Irnich *et al.*, “Spectrum sharing scenarios and resulting technical requirements for 5G systems," in *Proc. IEEE PIMRC*, pp. 127-132, Sep. 2013.

[61] W. Lehr and N. Jesuale, “Spectrum pooling for next generation public safety radio systems," in *Proc. IEEE DySPAN*, pp. 1-23, Oct. 2008.

[62] P. Si, H. Ji, F. Yu and V. Leung, “Optimal cooperative internetwork spectrum sharing for cognitive radio systems with spectrum pooling," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1760-1768, May 2010.

[63] G. Middleton *et al.*, “Inter-operator spectrum sharing in a broadband cellular network," in Proc. IEEE ISSSTA, pp. 376-380, Aug. 2006.

[64] L. Anchora *et al.*, “Capacity gains due to orthogonal spectrum sharing in multi-operator LTE cellular networks," in Proc. ISWCS, pp. 286-290, Aug. 2012.

[65] E. A. Jorswieck *et al.*, “Spectrum sharing improves the network efficiency for cellular operators," *IEEE Commun. Mag.*, vol. 52, no. 3, pp. 129-136, Mar. 2014.

[66] J. E. Suris *et al.*, “Cooperative game theory for distributed spectrum sharing," in *Proc. IEEE ICC*, pp. 5282-5287, Jun. 2007.

[67] S. Hailu *et al.*, “One-shot games for spectrum sharing among co-located radio access networks,” in *Proc. IEEE ICCS*, pp. 61-66, Nov. 2014.

[68] R. Etkin, A. Parekh and D. Tse, “Spectrum sharing for unlicensed bands," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 517-528, Apr. 2007.

[69] Y. Wu *et al.*, “Repeated open spectrum sharing game with cheat-proof strategies," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1922-1933, Nov. 2009

[70] H. Kamal, M. Coupechoux and P. Godlewski, “Inter-operator spectrum sharing for cellular networks using game theory," in *Proc. IEEE PIMRC*, pp. 425-429, Sep. 2009.

[71] W. S. Neilson, “The economics of favors," *J. Econ. Behavior Organization*, vol. 39, no. 4, pp 387-397, Jul. 1999.

[72] B. Singh, K. Koufos and O. Tirkkonen, “Coordination protocol for inter-operator spectrum sharing based on spectrum usage favors," in *Proc. EuCNC*, Jun. 2014.

[73] B. Singh, “Repeated games for inter-operator spectrum sharing," M.Sc. thesis, School of Electrical Engineering, Aalto University, Espoo, Jun. 2014.

[74] M. J. Osborne, *An Introduction to Game Theory*, Oxford: Oxford University Press, 2003.

[75] R. Axelrod and W. D. Hamilton, “The evolution of cooperation," *AAAS Sci.*, vol. 211, no. 4489, pp. 1390-1396, Mat. 1981.

[76] S. F. Cheng *et al.*, “Notes on equilibria in symmetric games," in *Proc. Workshop GTDT*, pp. 71-78, Aug. 2004.

[77] B. Singh, K. Koufos and O. Tirkkonen, “Co-primary inter-operator spectrum sharing using repeated games," in *Proc. IEEE ICCS*, pp. 67-71, Nov. 2014.

[78] G. Fettweis *et al.*, “The tactile internet," ITU-T, Tech. Watch Rep., Aug. 2014.

[79] P. Popovski, “Ultra-reliable communication in 5G wireless systems," in *Proc. Int. Conf. 5GU*, pp. 146-151, Nov. 2014.

[80] ETSI, “Speech and multimedia Transmission Quality (STQ); QoS of connections from current technologies to LTE for delay sensitive applications," ETSI TR 103 122 V1.1.1, Tech. Rep., Nov. 2011.

[81] L. Shi and K. W. Sung, “Spectrum requirement for vehicle-to-vehicle communication for traffic safety," in *Proc. IEEE VTC Spring*, pp. 1-5, May 2014.

[82] D. Xu *et al.*, “Elastic service availability: utility framework and optimal provisioning," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 6, pp. 55-65, Aug. 2008.

[83] M. Simsek *et al.*, “5G-enabled tactile internet," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 460-473, Mar. 2016.

[84] H. Shariatmadari, S. Iraji and R. Jäntti, “Analysis of transmission methods for ultra-reliable communications," in *Proc. IEEE PIMRC Workshop*, pp. 2303-2308, Aug. 2015.

[85] U. Oruthota, F. Ahmed and O. Tirkkonen, “Ultra-reliable link adaptation for downlink MISO transmission in 5G cellular networks," *Information*, vol. 7, no. 1, pp. 1-14, Mar. 2016.

[86] N. Brahmi *et al.*, “Deployment Strategies for ultra-reliable and low-latency communication in factory automation," in *Proc. IEEE GLOBECOM Workshop*, pp. 1-6, Dec. 2015.

[87] O. N. C. Yilmaz *et al.*, “Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *Proc. IEEE ICC Workshop*, pp. 1190-1195, Jun. 2015.

[88] N. A. Johansson *et al.*, “Radio access for ultra-reliable and low-latency 5G communications," in *Proc. IEEE ICC Workshop*, pp. 1184-1189, Jun. 2015.

[89] A.-R. Pitaval *et al.*, “Full-duplex self-backhauling for small-cell 5G networks," *IEEE Wireless Commun.*, vol. 22, no. 5, pp. 83-89, Oct. 2015.

[90] H. H. Tan and H. Wang, “Performance of multiple parallel slotted ALOHA channels," in *Proc. INFOCOM*, pp. 931-940, Mar. 1987.

[91] M. Tavana, V. S.-Mansouri and V. W. S. Wong, “Congestion control for bursty M2M traffic in LTE networks," in *Proc. IEEE ICC*, pp. 5815-5820, Jun. 2015.

[92] E. Paolini *et al.*, “Coded random access: Applying codes on graphs to design random access protocols," *IEEE Commun. Mag.*, vol. 53, no. 6, pp. 144-150, Jun. 2015.

[93] M. Ghanbarinejad and C. Schlegel, “Irregular repetition slotted ALOHA with multiuser detection," in *Proc. WONS*, pp. 201-205, Aug. 2013.

[94] C. Boyd, R. Vehkalahti and O. Tirkkonen, “Interference cancelling codes for ultra-reliable random access," *Int. J. Wireless Inf. Netw.*, vol. 25, no. 4, pp. 422–433, Jul. 2018.

[95] B. Singh *et al.*, “Identification of colliding UE for grant-free transmission of data and scheduling request based on resource sub-pool assignments for wireless networks," U.S. Patent 10383140, Aug. 2019.

[96] O. Tirkkonen *et al.*, “Multihop relaying for local area access," in *Proc. IEEE VTC Spring*, pp. 1-5, Jun. 2013.

[97] D. W. K. Ng, E. S. Lo and R. Schober, “Dynamic resource allocation in MIMO-OFDM systems with full-duplex and hybrid relaying," *IEEE Trans. Commun.*, vol. 60, no. 5, pp. 1291-1304, May 2012.

[98] Nokia and Alcatel-Lucent Shanghai Bell, “Contention-based grant-free transmission for URLLC," 3GPP TSG-RAN WG1 Meeting NR#1, Jan. 2017.

[99] B. Singh and M. Gerami, “Configured grant for semi-deterministic traffic for ultra-reliable and low latency communications," in *Proc. 6G Wireless Summit*, pp. 1-2, Mar. 2020.

[100] B. Singh *et al.*, “Grant-free transmissions for ultra-reliable and low latency uplink communications," in *Proc. EuCNC*, pp. 1-2, Jun. 2018.

[101] G. Choudhury and S. Rappaport, “Diversity ALOHA-A random access scheme for satellite communication," *IEEE Trans. Commun.*, vol. 31, no. 3, pp. 450-457, Mar. 1983.

[102] 3GPP, “Release 17” Available: `https://www.3gpp.org/release-17`. [Accessed: Jun. 21, 2021].

[103] D. Johnson, “With 5G rollout lagging, research looks ahead to 6G," Jul. 2020. [Online]. Available: `https://spectrum.ieee.org/tech-talk/telecom/wireless/with-5g-rollout-lagging-research-looks-ahead-to-6g`. [Accessed: Jun. 21, 2021].

[104] H. Viswanathan and P. E. Mogensen, “Communications in the 6G era," *IEEE Access*, vol. 8, pp. 57063-57074, Mar. 2020.

[105] 3GPP, “Study on enhancement for data collection for NR and EN-DC," TR 37.817 V0.1.0, Jan. 2021.

[106] 3GPP, “Study on traffic characteristics and performance requirements for AI/ML model transfer in 5GS," TR 22.874 V2.0.0, Jun. 2021.

Aalto-DD 169/2021

ISBN 978-952-64-0614-5 (printed)
ISBN 978-952-64-0615-2 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Electrical Engineering
Department of Communications and Networking
www.aalto.fi

BUSINESS + ECONOMY

ART + DESIGN + ARCHITECTURE

SCIENCE + TECHNOLOGY

CROSSOVER

DOCTORAL DISSERTATIONS